

# nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE



## Down to the bone

*Small-angle X-ray scattering and computer tomography combine to visualize hard tissue* **PAGES 308, 349 & 353**

SPACE-TIME GEOMETRY

### A BRUSH WITH GRAVITY

*Quantum entanglement versus relativity*

PAGE 290

GENERAL RELATIVITY

### EINSTEIN'S HELPERS

*Rivals and collaborators who made the theory possible*

PAGE 298

MATHEMATICS

### LOGICIAN AT PLAY

*Lewis Carroll's adventures in Wonderland and beyond*

PAGE 302

NATURE.COM/NATURE

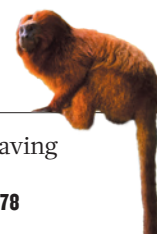
19 November 2015 £10

Vol. 527, No. 7578





# THIS WEEK



## EDITORIALS

**SPORT** Science is no antidote to doping if the culture condones it **p.276**

**WORLD VIEW** Time to make universities more democratic **p.277**

**CONSERVATION** Saving species from extinction **p.278**

## Research for all

*Numbers on racial bias in research grants awarded by the US National Institutes of Health show that science has more to learn about inclusiveness.*

Data released this week paint a long-term picture of the racial disparity in grants funded by the US National Institutes of Health (NIH), and show that for nearly 30 years, applicants from minorities have been less successful than white and mixed-race applicants in receiving funding (see page 286). The data, obtained by researchers through a Freedom of Information Act (FOIA) request to the agency, extend previous findings that showed racial disparities in NIH grants between 2000 and 2006. Those results had already led the agency to dedicate hundreds of millions of dollars' worth of grants and programmes to try to rectify funding disparities.

These disparities are not only unjust; they harm scientific and medical progress by shutting off funding to deserving scientists, and ultimately harm society and patients who would otherwise benefit from these scientists' ideas.

The big task now is to determine why racial funding disparities arise, and how to erase them. Researchers who have studied such disparities in science say that bias probably plays out in subtle and unsubtle ways in the grant-review process itself.

In 2011, a team led by Raynard Kington, president of Grinnell College in Iowa, published a landmark paper reporting that black grant applicants were about two-thirds as likely as whites to be funded during 2000–06, even once factors such as publication history and training were taken into account (D. K. Ginther *et al. Science* **333**, 1015–1019; 2011).

Initially, the same study also found funding disparities between Asians and whites. But when the authors controlled for nativity — that is, when they attempted to distinguish Asians who were born and educated in the United States from those who had immigrated after receiving some or all of their scientific training — they found that there was no disparity in funding between whites and Asians who were US citizens when they received their PhDs.

Kington says that this finding shows that bias can work in more complex ways than along strict racial lines: for instance, in favour of native-born and against foreign-born grant applicants. Countering this effect requires changes to the peer-review process, but also action throughout scientists' education, training and career trajectories.

That will need measures to counter negative bias — the tendency to have stereotyped or unfavourable opinions about people belonging

**“Bias can work in more complex ways than along strict racial lines.”**

to particular groups — but it will also require strategies that take into account positive bias, which is the tendency to support, like and believe in people who are similar to, or have similar experiences to, oneself. Much of the disparity among funded researchers is thought to involve factors such as whether

grant applicants trained at or work at the same institutions, study the same research questions and have published as much or in the same journals as peer reviewers.

The NIH is working on some aspects of the issue — for instance, its National Research Mentoring Network aims to foster diversity through mentoring. Pulmonologist Esteban Burchard and epidemiologist Sam Oh at the University of California, San Francisco, who requested the NIH FOIA data, support ideas that would require NIH grant reviewers to score grant applications on factors such as how the researchers plan to recruit diverse populations and how well the applications reflect the racial and ethnic make-up of the country. Another idea is to provide an administrative supplement to funded grants for the discovery of racial and ethnic differences in medicine.

Attacking such disparities — both in the ranks of science itself and at the level of the type of science funded — is a powerful idea. It deserves serious consideration as the NIH works to make sure that the research it funds is truly representative of the medical needs of the United States. ■

## Defensive drives

*Researchers exploring ways to genetically alter wild populations are wise to air their plans.*

Often, scientists in a fast-moving field try to keep a tight lid on their work until it is published. But the authors of a paper published in *Nature Biotechnology* this week have been unusually chatty about their work, broadcasting their results over the past year and airing their plans for further research.

Because the work would make it possible to modify the genetics of entire populations of organisms, it raises a host of ethical and safety

questions. The researchers consider it wise to prepare their colleagues and the public for the results to come, and to solicit suggestions from the community about how to execute such experiments safely. The technique in question is an engineered 'gene drive' — a system that can spread a mutation through a population much faster than normal. The practice could wipe out some insect-borne diseases, including malaria. But an accidental release could trigger unintended, ecosystem-wide consequences. As such, research that involves gene drives must be handled with utmost care.

The paper, published on 16 November, could ease concerns about accidental releases (J. E. DiCarlo *et al. Nature Biotechnol.* <http://doi.org/89h>; 2015). A team of researchers from Harvard Medical School in Boston, Massachusetts, has demonstrated that gene drives can be engineered that will work in laboratory strains of yeast (*Saccharomyces cerevisiae*), but that are unlikely to function in wild populations. And

if one did escape, the team showed that the mutation can be undone by setting loose a second drive to 'overwrite' it. The results suggest that careful planning can reduce the risks while allowing gene-drive research to continue.

The concept of a gene drive is decades old, but the technique's application was hindered until the discovery a few years ago of a simple and versatile genome-editing system called CRISPR-Cas9. This system allows researchers to alter genomes with unprecedented precision and to engineer the fundamental components of a gene drive that transmit a copy of the edited sequence to nearly all offspring.

And because the CRISPR-Cas9 system is relatively easy to work with, the technology is now available to more laboratories than ever before. This is both a boon and a concern: it arms more great minds with a tool that could address serious public-health and environmental problems, but it also increases the chance that a laboratory might enter the field naive to the necessary safety precautions. This has understandably raised some safety concerns, and the US National Academies of Sciences, Engineering, and Medicine, for example, has convened a committee to evaluate uses of gene drives.

With this in mind, the Harvard researchers have been careful to announce their experimental plans before they carry them out. The experiments published this week were alluded to in previous publications outlining safety precautions that could be taken. The authors gathered feedback from the community, and used this to boost the safety of their own experiments. They were also careful

to develop these safeguards before carrying out key laboratory experiments to explore the use of gene drives against Lyme disease, which is transmitted to humans through ticks, and schistosomiasis, a scourge carried by parasitic worms and most often found in Africa. All of these experiments have been discussed openly, before they were carried out.

Such openness is not standard practice. Scientific experiments are often subject to approval by institutional safety committees and funding agencies, but these discussions tend to be carried out behind closed doors. The public is sometimes surprised by what emerges. Witness the reaction earlier this year, when researchers announced that they had used CRISPR-Cas9 to edit the genomes of human embryos (see *Nature* **520**, 593–595; 2015). About three years ago, a charged debate and research moratorium ensued when news broke that researchers intended to publish results showing how they had engineered the H5N1 influenza virus to make it more transmissible.

The Harvard gene-drive researchers have learnt from these debacles, and recognized the need to alert the wider community to their plans so that discussions can take place, concerns can be aired and suggestions offered from all corners before the work is done. Scientists should watch closely to see whether this approach could serve as a template for other teams that take on the challenge of working in controversial fields. ■

**“Research involving gene drives must be handled with utmost care.”**

# Dope rules

*Science is beside the point when an entrenched culture in a sport supports scoundrels.*

This month, the World Anti-Doping Agency (WADA) dropped a bombshell on the athletics world. In a scathing 335-page report issued on 9 November, the independent international agency alleged the existence of a far-reaching doping programme in Russian track and field that implicated government officials, sporting federations, coaches, athletes, scientists and doctors. As a result of Russia's widespread and institutionalized doping programme, the 2012 Olympic Games in London were effectively “sabotaged”, WADA concluded.

Russian athletes and coaches have subsequently been suspended from international track-and-field competitions by the International Association of Athletics Federations, whose laissez-faire policies over the years contributed to the scandal, according to WADA. There is a real possibility that Russian athletes will be banned from the 2016 Olympic Games in Rio de Janeiro, Brazil.

Science, some argue, can lead the way in achieving clean, or at least cleaner, sport. In the face of improved tests to detect doping, it will become harder for athletes to ingest performance-enhancing drugs without getting caught. Many say that the answer might lie with the biological passport, which looks for changes in blood chemistry from an individual's 'baseline' profile that may be indicative of doping. Suspect blood profiles have been used to nab cheats in professional cycling and endurance sports such as biathlons. And there have been claims from scientists at cycling's international federation, the UCI — itself subject to allegations of misconduct — that rampant blood doping became less common in the pro peloton (the elite professional cycling circuit) after biological passports were introduced.

There is no doubt that science can play a major part in anti-doping efforts. But this can only happen once a governance system is in place that has a genuine interest in clean sport. A series of doping scandals

has shown that science is useless at catching cheats in a culture that doesn't really want to catch them — and in many cases is being used to help them.

Sophisticated biological passports are futile in a culture that encourages a leading anti-doping scientist to destroy blood and urine samples by the hundreds, while extorting money from athletes to do so. This is what WADA's report alleges of Grigory Rodchenkov, the former head of Russia's leading anti-doping lab. And the UCI looked the other way as Lance Armstrong doped his way to seven successive (and now rescinded) Tour de France victories, according to a report issued this year by an independent commission appointed by the UCI.

**“Sports that haven't been roiled by doping scandals may not be looking hard enough.”**

Sports that haven't been roiled by doping scandals may not be looking hard enough. Financial incentives such as corporate sponsorships, broadcasting rights and merchandizing have the potential to discourage strong and independent anti-doping programmes. Just look at professional cycling, which has been forced to confront its massive doping problems. The sport's popularity has suffered, and spectacular performances such as those of Britain's Chris Froome now raise just as many suspicions as celebrations. Just imagine if, during an international tennis tournament, a 150-m.p.h. serve raised eyebrows rather than awe.

There will be calls for more and better anti-doping tests in the run-up to next year's summer Olympics in Rio and in other high-profile competitions. These are genuinely needed, because dopers tend to be a step ahead. Officials will hail their cutting-edge laboratories full of gleaming mass spectrometers and haemocytometers, and brag about how many urine and blood tests these can process each day — never mind that savvy athletes tend to dope out of competition and in tiny doses that are nearly impossible to detect.

If past attitudes are anything to go by, we can expect officials to hide behind science, while doing little to root out the fundamental problems that allow systemic sports doping to thrive. As Russia's doping scandal shows us, it is much easier to change a test tube than it is to change a culture. ■

➔ **NATURE.COM**  
To comment online,  
click on Editorials at:  
[go.nature.com/xhunqv](http://go.nature.com/xhunqv)





## Time to cry out for academic freedom

*Giving staff and students a say in how institutions are run would strengthen governance and clip the wings of administrators, argues Colin Macilwain.*

It was Clark Kerr, a former chancellor of the University of California, Berkeley, who most memorably defined the role of a university administrator: to arrange parking for the staff, sex for the students and sports for the alumni.

Kerr's throwaway line contained a kernel of truth: university administration is a necessary evil. Students and academics are the heart and soul of a university, and do its real work. The administrators, as his definition implies, merely facilitate. Well, these days, you would hardly know it. On campuses across the world, managerialism is on the march. The ancient power struggle between academics and administrators has lurched decisively in favour of the latter.

Nowhere is this more true than in the United Kingdom, where reforms instigated by former prime minister Margaret Thatcher 30 years ago have enabled vice-chancellors to extend their grip over matters that were once controlled by academics, such as what subjects to teach, which kinds of grant to chase and criteria for hiring staff. By shifting decision-making to committees dominated by their close allies, many vice-chancellors now operate as if they were chief executives.

This managerial approach, and the growing reach and expanse of administrative staff that has accompanied it, is gathering pace worldwide. That is largely because British and US universities dominate international university league tables, and many countries' higher-education policies seek to emulate their model. Germany's Excellence Initiative, for example, has selected a small number of promising institutions and given them the money to build up stronger central administrations.

I am not convinced that any of this is in the interests of students, teachers or the wider communities that universities are supposed to serve. I am a UK resident, but share continental suspicions of league tables, most of which originated in the English-speaking world and reflect the strengths of institutions there. At Europe-wide meetings, I often empathize with the audience's evident bemusement at the self-regard with which British and US speakers comport themselves. I fear that, left unchecked, the high-handed behaviour of some vice-chancellors will do to UK universities what the bearers of the same corporate outlook have done to our once-respected high-street banks. Just as local bank managers are a thing of the past, stubbornly independent academics can be flushed out by aggressive administrators, with their management speak and freshly minted data on research performance.

Most of the resistance to these changes has come from the arts and humanities. Critics such as literary scholar Stefan Collini have lambasted what they regard as the destruction of the UK university system in the *London Review of Books* and elsewhere. Scientists have been largely favoured by the new regime,

and, perhaps as a result, their leaders have been mute on the issue.

That cannot last. Research funding in the United Kingdom has been frozen for five years and, after next week's government spending review, will probably start to decline. About to enter the fray is the proposed Teaching Excellence Framework. This is modelled on the Research Excellence Framework for evaluating university research, which has dominated British higher education since 1986 and has been widely adopted overseas. The pressures placed on academics by often-bogus metrics continue to mount.

Here in Scotland, and rather against the tide, an effort is under way to give staff and students more influence over how universities are run. A higher-education governance bill proposed by the devolved government (see [go.nature.com/qqphw](http://go.nature.com/qqphw)) would mandate elections

of those wishing to chair university governing bodies, and would require the inclusion of staff and student unions on major committees. The bill takes its cue from a 2012 report led by Ferdinand von Prondzynski, principal and vice-chancellor of Robert Gordon University in Aberdeen, and includes measures to guarantee academic freedom under the law (see [go.nature.com/jvoma5](http://go.nature.com/jvoma5)).

The modest reforms in the bill are being fought tooth and nail by other university principals. Their association, Universities Scotland, has worked hard to create the impression that academics are happy with the current governance arrangements. It has argued that parts of the bill threaten the institutions' autonomy and would endanger their charitable status. It has even won the backing of Edward Snowden, the celebrated US whistleblower and student-elected rector of the University of Glasgow, who has tweeted opposition to the bill from his bunker in Moscow. But the proposal merely takes the process that got Snowden his position at Glasgow and extends it to other, newer institutions.

The Scottish National Party government and the opposition Labour Party are both likely to support passage of the bill, modified somewhat to protect institutions' autonomy. So universities in this corner of the world, at least, are set to become a touch more democratic. I'd like to see that movement spread to other places, where governments have been neglecting the concepts of democracy and academic freedom in their relentless push to make universities more businesslike. The cult of the chief executive has already permeated almost every corner of our society. But universities are not corporations. They are, by definition, self-organized bodies of academics. Student and staff representation on key university committees will remind administrators of this fact, call them to account — and, in the long run, strengthen university governance. ■

**Colin Macilwain** writes about science policy from Edinburgh, UK.  
e-mail: [cfmworldview@googlemail.com](mailto:cfmworldview@googlemail.com)

GOVERNMENTS  
HAVE NEGLECTED  
**DEMOCRACY**  
AND ACADEMIC  
FREEDOM IN THEIR  
PUSH TO MAKE  
UNIVERSITIES MORE  
**BUSINESSLIKE.**

➔ **NATURE.COM**  
Discuss this article  
online at:  
[go.nature.com/cozh8d](http://go.nature.com/cozh8d)

# RESEARCH HIGHLIGHTS

Selections from the  
scientific literature

## GENETIC ENGINEERING

### Boosting 'gene drive' safety

Researchers have developed a way to reduce the risks of a method that genetically engineers entire populations with unprecedented speed.

'Gene drives' are genetic changes, based on inserting parts of the CRISPR-Cas9 genome-editing system into a host genome, that spread through a population more rapidly than do normal mutations. Gene drives could be used to wipe out disease-carrying insects, for example, but could also spread uncontrollably in an ecosystem.

To reduce this risk, Kevin Esvelt and George Church of Harvard Medical School in Boston, Massachusetts, and their team inserted the bacterial DNA-cutting Cas9 enzyme into a piece of DNA external to the *Saccharomyces cerevisiae* yeast genome, and put the guide RNAs for directing Cas9 to a specific DNA sequence into the genome. This separation ensured that the gene drive would not spread exponentially if the strain was released into the wild.

*Nature Biotech.* <http://doi.org/89h> (2015)

## ANIMAL BEHAVIOUR

### Polarized light as a secret signal

Some crustaceans can detect polarized light, using it as a covert signal that is



invisible to predators.

Yakir Luc Gagnon at the University of Queensland in Brisbane, Australia, found that the bodies of mantis shrimps (*Gonodactylaceus falcatus*; pictured) reflect a distinctive pattern of circular polarization (pictured in red) that is visible only to other shrimps. When presented with different burrows in the laboratory, mantis shrimps avoided or delayed entering those that were lit with circularly polarized light compared with those under unpolarized light. This suggests that the shrimps use polarized light cues to sense whether potential burrows are occupied.

In another study, Martin

How at the University of Bristol, UK, found that male fiddler crabs (*Uca stenodactylus*) detected polarized targets in the wild from farther away than non-polarized ones. The animals' sensitivity to polarized light could be boosting the visual contrast between crabs and their mudflat habitat. *Curr. Biol.* <http://doi.org/89c> (2015); <http://doi.org/89d> (2015)

## MICROBIOLOGY

### Antibiotics make MRSA worse

Antibiotics could help a drug-resistant pathogen to worsen inflammation.

Whereas national-level legislation did not necessarily lead to good outcomes, intensity of local law enforcement did. Moreover, reducing the threats to animals — such as habitat loss and hunting — was crucial for long-term survival.

No link was found between the outcomes of recovery programmes and biological factors such as body mass and habitat type, suggesting that well-designed conservation programmes should work across different species.

*Conserv. Biol.* <http://doi.org/87v> (2015)

Methicillin-resistant *Staphylococcus aureus* (MRSA) resists most  $\beta$ -lactam antibiotics by acquiring a protein that modifies the cell wall. David Underhill, George Liu and their team at Cedars-Sinai Medical Center in Los Angeles, California, thought that this modification might also boost the production of inflammatory molecules called cytokines in the host. They exposed mouse and human immune cells to MRSA and found that the host cells made higher levels of a cytokine called IL-1 $\beta$  when MRSA had been grown in the presence of  $\beta$ -lactams. In mice, treatment with a  $\beta$ -lactam caused more immune cells to flood the site



## CONSERVATION

### How to save a species

Common factors such as law enforcement contribute to the success or failure of species-recovery programmes, suggesting that conservation lessons could be generalized across different populations or species.

Jennifer Crees at the Zoological Society of London and her colleagues analysed 48 mammalian conservation programmes, ranging from the successful protection of the golden lion tamarin (pictured) to the failed attempt to save the Yangtze River dolphin.

KIKE CALVO/NATIONAL GEOGRAPHIC/GETTY

WITH PERMISSION FROM ELSEVIER



of an MRSA skin infection, resulting in more inflammation and larger abscesses than in MRSA-infected mice that were not treated with the antibiotic.

MRSA infections are still sometimes treated with  $\beta$ -lactams, but these should be used with other antibiotic classes, the authors write.

*Cell Host Microbe* 18, 604–612 (2015)

## MUSCLE BIOLOGY

## Dog saved from muscular disease

A golden retriever with the mutation for Duchenne muscular dystrophy was found to have working muscles because of a compensatory mutation in another gene.

The dog, Ringo, was bred by researchers to have the mutated version of a protein called dystrophin, but he still had normal muscles. Louis Kunkel at Boston Children's Hospital in Massachusetts, Mayana Zatz at the University of São Paulo in Brazil and their team analysed the genomes of Ringo and one of his male offspring that also had the mutation and normal muscles. They identified a separate mutation in a development gene, *Jagged1*, that resulted in higher levels of *Jagged1* in Ringo and his son than in 31 affected dogs.

This mutation may compensate for the muscle-regeneration problems caused by a lack of dystrophin, the authors suggest.

*Cell* <http://doi.org/87s> (2015)

## BIOTECHNOLOGY

## Adult cells edited and reprogrammed

A one-step procedure can correct genetic mutations in body cells and reprogram them into stem cells.

Stem cells derived from patients' tissues could generate replacement tissue that is not rejected by the immune system. Current methods have stem-cell yields of only 0.5–0.9%, and require extra steps to correct any mutations. Sara

Howden, now at the Murdoch Children's Research Institute in Parkville, Australia, and her team introduced into cells a mix of genes that induce stem-cell formation and encode the components of the CRISPR–Cas9 gene-editing system.

They targeted mutations in cells from two people — an adult with a degenerative retinal disease and an infant with severe combined immunodeficiency — and made stem cells without the defect. The method produced stem cells with 5–8% efficiency. *Stem Cell Rep.* <http://doi.org/87t> (2015)

## NATURAL HISTORY

## Selenium linked to mass extinctions

Plummeting ocean reserves of selenium could have played a part in past mass extinctions.

Selenium and other trace elements help certain enzymes to function and perform other essential biochemical duties in organisms. John Long at Flinders University in Adelaide, Australia, and his team estimated ocean selenium levels over the past 560 million years by analysing it in marine pyrite samples. Selenium concentrations fluctuated drastically, but sharp drops coincided with several mass extinctions — including one at the end of the Triassic 200 million years ago.

Crashes in selenium levels may have acted in concert with changes in oxygen and carbon cycles to drive mass extinctions, the authors say. *Gondwana Res.* <http://doi.org/834> (2015)

## MATERIALS

## Super-thin superconductor

A layer of niobium diselenide ( $\text{NbSe}_2$ ) just a few atoms thick can conduct electricity with zero resistance.

Most 3D superconducting materials lose this ability once they are in their 2D form. Miguel Ugeda at the

## SOCIAL SELECTION

Popular topics  
on social media

## How to judge scientists' strengths

When the director of a research institute asked his Twitter followers for a practical way to dig out promising candidates from the hundreds of applications sitting on his desk, they responded in spades. Ewan Birney, co-director of the EMBL European Bioinformatics Institute in Hinxton, UK, tweeted that he was procrastinating over how to shortlist the applications, which together listed around 2,500 research papers. (The process is ongoing, so Birney would not say exactly what the researchers were applying for.) He tweeted: "I get \*genuinely\* stuck here. If I am not going to use journal title as a proxy for quality, what do I do?"

➔ NATURE.COM

For more on

popular papers:

[go.nature.com/pim8hf](http://go.nature.com/pim8hf)

Yoav Gilad, a geneticist at the University of Chicago, Illinois, tweeted: "Read the abstracts. Read the papers. Yes, if it means 2500 papers, then get a larger committee."

nanoGUNE research centre in San Sebastian, Spain, Michael Crommie at the University of California, Berkeley, and their colleagues studied the behaviour of electrons in a single layer of  $\text{NbSe}_2$ , grown on a bilayer of atom-thick carbon.

As the team lowered the temperature to below  $-271^\circ\text{C}$ , the material's resistance fell to zero. The authors say that the results confirm that  $\text{NbSe}_2$  is a true 2D superconductor, a class of materials that could one day be used in tiny quantum computers and other devices.

The team also saw ripples in electron density while the material was superconducting — an effect that some theories predict should not be possible. *Nature Phys.* <http://doi.org/89g> (2015)

## ENVIRONMENTAL SCIENCE

## Chemicals hinder oil-eating microbes

Chemical dispersants added to spilled oil from the 2010 Deepwater Horizon disaster in the Gulf of Mexico (pictured) may have made little difference to the rates at which microbes broke down the oil.

The dispersants broke up the oil into smaller droplets to help sea-dwelling microbes to degrade it further. To study the chemicals' effect on the



microbes, Samantha Joye at the University of Georgia in Athens and her colleagues created bottled mixtures of sea water, oil and dispersants that simulated environmental conditions during the spill.

Mixtures of oil and sea water were dominated by *Marinobacter* species, which can degrade a wide range of hydrocarbons. But these populations dropped when dispersant was added, whereas *Colwellia*, which degrades dispersants, increased in abundance. Adding dispersants did not seem to change the rate at which hydrocarbons were broken down in the bottled samples. *Proc. Natl Acad. Sci. USA* <http://doi.org/89f> (2015)

➔ NATURE.COM

For the latest research published by Nature visit:

[www.nature.com/latestresearch](http://www.nature.com/latestresearch)

# SEVEN DAYS

The news in brief

## BUSINESS

### Ocata buyout

Astellas Pharma in Tokyo will pay US\$379 million for Ocata Therapeutics of Marlborough, Massachusetts. Formerly called Advanced Cell Technology, Ocata has struggled financially but has continued to develop treatments in which human embryonic stem cells are coaxed into becoming retinal cells. The company has used the cells to treat two types of degenerative blindness in small-scale clinical trials. In the United States, limiting trials to a few participants would hamper speedy commercialization, but Japan has a fast-track approval system that allows commercialization of stem-cell treatments after studies on a small number of people.

## EVENTS

### Paris talks go ahead

International climate talks in Paris will go ahead despite the 13 November terrorist attacks that killed at least 129 people in the French capital. The climate conference will be held, with tightened security, because it is an “essential meeting for humanity”, French Prime Minister Manuel Valls

## NUMBER CRUNCH

# 8,690

The wind speed in kilometres per hour on HD 189733 b, a ‘hot Jupiter’ exoplanet 19.3 parsecs away from Earth — and the first weather data from a planet outside our Solar System.

Source: Louden, T. & Wheatley, P. J. Preprint at <http://arxiv.org/abs/1511.03689> (2015).



IAC/UA SPACE AGENCY/NASA/ESA

## Space junk splashes down safely

A chunk of space debris re-entered Earth’s atmosphere on 13 November. The fragment was too small to hurt anyone but just the right size to help scientists to practise tracking an incoming asteroid. Researchers on a chartered jet filmed the debris, which may have fallen

off a lunar spacecraft, as it disintegrated above the Indian Ocean near Sri Lanka. NASA astronomer Peter Jenniskens says the successful campaign proves that it is possible to gather data about an object targeting the planet, even with short notice.

said on 14 November. Some 40,000 participants will gather for the United Nations climate summit from 30 November to 11 December. Almost 120 government leaders will attend the meeting, which it is hoped will produce a global climate deal.

### Dark-matter hunt

The world’s most sensitive detector for dark matter was inaugurated on 11 November at the Gran Sasso National Laboratory, run by Italy’s National Institute for Nuclear Physics. Dark matter is thought to make up 85% of matter in the Universe. The experiment, called XENON1T, will monitor 3.5 tonnes of liquid xenon, to try to detect the tiny amount of energy that is given off when dark matter interacts with atoms of ordinary matter. The collaboration

involves 125 scientists, and the experiment is expected to start collecting data by the end of March 2016.

### Nuclear burial

Finland’s government approved the construction of a deep underground facility to permanently store spent nuclear fuel on 12 November. Minister of economic affairs Olli Rehn said the move was a world first. The repository will dispose of up to 6,500 tonnes of uranium — high-level waste produced by nuclear-power facilities — by packing it into copper canisters and burying these in a clay buffer 400 metres underground. The local government has already given its go-ahead for the facility, which will be on Olkiluoto island off Finland’s west coast and is due to open around 2023.

## POLICY

### Pesticide risk

The world’s most widely used herbicide, glyphosate, is unlikely to pose a cancer risk to humans, according to a report published on 12 November by the European Food Safety Authority (EFSA). In its report, the agency set limits on how much glyphosate a person may safely ingest in a short period of time. EFSA’s finding comes nearly eight months after the World Health Organization’s International Agency for Research on Cancer said that glyphosate probably does cause cancer in humans. See [go.nature.com/mb8b4l](http://go.nature.com/mb8b4l) for more.

### Space mining is go

On 10 November the US Senate passed the Space Act of 2015, allowing US citizens the



rights to any materials that they gather from asteroids or other space-based resources. However, space miners will also have to comply with the 1966 Outer Space Treaty, an international agreement that states: "outer space is not subject to national appropriation by claim of sovereignty". The act also extends the use of the International Space Station from 2020 to at least 2024.

## Pipeline questions

In a letter to transport minister Marc Garneau, Canadian Prime Minister Justin Trudeau on 13 November called for a moratorium on crude-oil-tanker traffic along the north coast of British Columbia. The move raises questions about a pipeline project by the Calgary-based energy-delivery firm Enbridge to carry oil from Alberta's tar sands to the coast for shipment. Environmentalists said that a moratorium would effectively halt the pipeline, but Enbridge said that it still hopes to discuss the plan with the prime minister.

## Reef protected

Laws passed on 12 November in Queensland, Australia, will protect the Great Barrier Reef (pictured) from port development, the state's development minister said. The laws ban disposal at sea of



material dredged from ports in the region, and stop any new ports being developed in the reef World Heritage Area. They form part of commitments made by Australia to safeguard the reef after the United Nations Educational, Scientific and Cultural Organization considered categorizing the coral zone as 'at risk'. Plans to dispose of dredging material near the reef have proved controversial in recent years, and conservation groups welcomed last week's legislation.

## Food rules

For the first time, crop farmers in the United States will have to answer to the Food and Drug Administration (FDA) as part of an effort to prevent food-borne illness. A set of

rules that the agency released on 13 November requires farmers to train their workers in proper hygiene, and to test crop-irrigation systems for pathogens, among other things. But the regulations are less stringent than a 2013 FDA proposal that farmers found too burdensome. Another of the rules creates a programme to allow auditors to assess imported food and the overseas facilities that produce it.

## FUNDING

## Olive aid

The European Commission has announced a €7-million (US\$7.5-million) call for proposals for research into *Xylella fastidiosa*, the aggressive plant pathogen that is destroying swathes

## COMING UP

### 19–20 NOVEMBER

The inaugural Meta-Research Innovation Center at Stanford (METRICS) conference convenes in California to discuss how biomedical scientists plan, conduct and communicate research. [go.nature.com/narepc](http://go.nature.com/narepc)

### 23–27 NOVEMBER

Ostend, Belgium, hosts the European Space Weather Week, a forum for space-weather forecasters and scientists. [www.stce.be/esww12](http://www.stce.be/esww12)

### 24–26 NOVEMBER

An international immunotherapy conference meets in Brisbane, Australia. [go.nature.com/235ing](http://go.nature.com/235ing)

of olive trees in the Puglia region of southern Italy. The call will focus on methods of detection and control. The outbreak, which has also reached some regions of France, is a major economic threat to the European Union, but has received little research funding so far. Italian regional and national governments have also promised €6 million for *X. fastidiosa* research.

## Chile budget boost

Chile's Congress was mulling a budget increase of 150 million pesos (US\$210,000) for the nation's research-funding agency as *Nature* went to press. The move followed street protests by researchers after the resignation of Francisco Brieva, director of the National Commission for Scientific and Technological Research. The body funds more than 3,000 researchers. See [go.nature.com/pbwtp8](http://go.nature.com/pbwtp8) for more.

## NATURE.COM

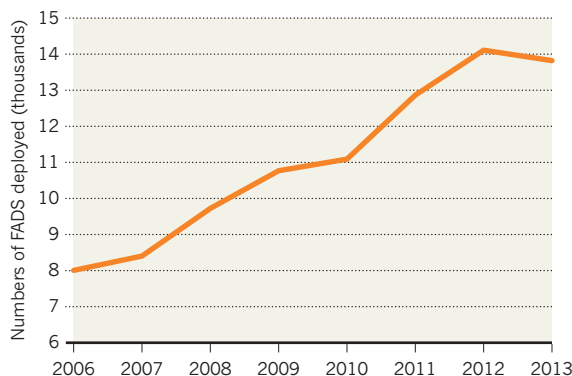
For daily news updates see: [www.nature.com/news](http://www.nature.com/news)

## TREND WATCH

The number of controversial fish aggregating devices (FADs) being used in the oceans is rising. Using data from tuna-fishing boats (see chart), a 6 November report from the Pew Charitable Trusts estimates that between 81,000 and 121,000 FADs were set adrift in 2013 — and their use is growing (see [go.nature.com/ngeubv](http://go.nature.com/ngeubv)). These FADs are free-floating, so fish and other animals shelter underneath and become easier to catch. But researchers warn that the devices encourage overfishing, and kill vulnerable species.

## 'FISH AGGREGATING DEVICE' USE ON THE RISE

Floating devices that could kill vulnerable species are being used increasingly, despite concerns from researchers and environmentalists, an analysis from tuna fishing boats shows.



# NEWS IN FOCUS

**SPACE** Prototype gravitational-wave observatory ready to launch **p.284**

**POLICY** US minorities less likely to get biomedical research grants **p.286**

**EXOPLANETS** Graphical guide to the next 20 years of discovery **p.288**



**PHYSICS** Quantum entanglement is the essence of space-time **p.290**

DOUG ALLAN/NATURE PICTURE LIBRARY



Beluga whales are among the species that are thought to use Earth's weak magnetic field for navigation.

## BIOCHEMISTRY

# Long-sought 'biocompass' discovery claimed

*Protein complex offers explanation for how animals sense Earth's magnetic pull.*

BY DAVID CYRANOSKI

In the cells of fruit flies, Chinese scientists say that they have found a biological compass needle: a rod-shaped complex of proteins that can align with Earth's weak magnetic field.

The biocompass — whose constituent proteins exist in related forms in other species, including humans — could explain a long-standing puzzle: how animals such as birds and insects sense magnetism. It might also become an invaluable tool for using magnetic fields to control cells, report researchers led by biophysicist Xie Can at Peking University in

Beijing, in a paper published on 16 November in *Nature Materials* (S. Qin *et al. Nature Mater.* <http://doi.org/89v>; 2015).

"It's an extraordinary paper," says Peter Hore, a biochemist at the University of Oxford, UK. But Xie's team has not shown that the complex actually behaves as a biocompass inside living cells, nor explained exactly how it senses magnetism. "It's either a very important paper or totally wrong. I strongly suspect the latter," says David Keays, a neuroscientist who studies magnetoreception at the Institute of Molecular Pathology in Vienna.

Many organisms — ranging from whales to butterflies, and termites to pigeons — use

Earth's magnetic field to navigate or orient themselves in space. But the molecular mechanism behind this ability, termed magneto-reception, is unclear.

Some researchers have pointed to magnetically sensitive proteins called 'cryptochromes', or 'Cry'. Fruit flies lacking the proteins lose their sensitivity to magnetic fields, for example. But the Cry proteins alone cannot act as a compass, says Xie, because they cannot sense the polarity (north-south orientation) of magnetic fields.

Others have suggested that iron-based minerals might be responsible. Magnetite, a form of iron oxide, has been found in the beak cells of homing pigeons. Yet studies suggest ►



► that magnetite plays no part in pigeon magnetoreception.

Xie says that he has found a protein in fruit flies that both binds to iron and interacts with Cry. Known as CG8198, it binds iron and sulfur atoms and is involved in fruit-fly circadian rhythms. Together with Cry, it forms a nanoscale 'needle': a rod-like core of CG8198 polymers with an outer layer of Cry proteins that twists around the core.

Using an electron microscope, Xie's team saw assemblies of these rods orienting themselves in a weak magnetic field in the same way as compass needles. Xie gave CG8198 the new name of MagR, for magnetic receptor.

The discovery offers scientists the prospect of using magnetic fields to control cells. Over the past decade, scientists have commandeered the light-sensing capacity of some proteins to manipulate neurons, usually by inserting a fibre-optic cable directly into the brain — a tool called optogenetics. But magnetosensing proteins have the advantage that they could be manipulated by magnetic fields outside the brain.

Zhang Sheng-jia, a neuroscientist at Tsinghua University in Beijing, claims to have already demonstrated this 'magnetogenetic' capability. In September, he provided a

surprise preview of Xie's work when he published a paper reporting use of the biocompass to manipulate neurons in worms (X. Long *et al. Sci. Bull.* <http://doi.org/883>; 2015). Xie and others complained that Zhang's early publication violated a collaboration agreement between the two researchers — the details of which are disputed — and asked for it to be retracted. In October, Zhang was fired from his university, a decision that he is contesting (see *Nature* <http://doi.org/882>; 2015).

**"If MagR is the real magnetoreceptor, I'll eat my hat."**

Xie says that in April, he submitted a Chinese patent application that includes the use of magnetogenetics and the protein's magnetic capacity to manipulate large molecules. He is also starting to look at the structure of MagR proteins in other animals, including humans. Variants in the human version of MagR might even relate to differences in people's sense of direction, he suggests.

#### SCPTICAL VOICES

Other scientists are not convinced that the biological needles function like compasses in living organisms. Xie's team has shown that MagR and Cry are produced in the same

cells in pigeon retinas — the birds' proposed magnetoreception centre — but MagR and Cry are found in many cells, says Keays. "With such a small amount of iron, one has to ask whether *in vivo*, at physiological temperatures, MagR is capable of possessing magnetic properties at all," he says. "If MagR is the real magnetoreceptor, I'll eat my hat."

Xie hopes that others will strengthen his case with further experiments, such as inactivating the gene for MagR in certain fruit-fly tissues to see whether it affects the animals' sense of direction. He published without doing this work, he says, because he just wanted to report the findings, which he has been working on for six years.

The lack of an exact mechanism for how the protein complex senses magnetism, or how any signal it sends might be processed by the brain, gives some researchers pause. MagR's biocompass activity might simply be the result of experimental contamination, says Michael Winklhofer, a magnetism specialist and Earth scientist at Ludwig Maximilian University of Munich in Germany. He is planning experiments to follow up on Xie's team's findings. If it holds up, says Winklhofer, then the discovery of MagR "appears to be a major step forward towards unravelling the molecular basis of magnetoreception". ■

#### PHYSICS

# Space test for long-awaited gravitational-wave detector

Europe's *LISA Pathfinder* spacecraft has two metal cubes at its heart, which it will attempt to isolate from every force except for gravity.

BY ELIZABETH GIBNEY

There is a lot riding on the *LISA Pathfinder* mission, an ambitious effort to test whether intricate technology designed to detect ripples in space-time can be deployed in space.

Scheduled to launch on 2 December, the spacecraft is a long-awaited test-drive for a future €1-billion (US\$1.1-billion) space observatory planned by the European Space Agency (ESA). The follow-up mission would track the largest objects in the Universe, including mergers between supermassive black holes and collisions between galaxies, by the space-time ripples that they create.

First predicted by Albert Einstein almost exactly 100 years ago as part of his general theory of relativity (see [nature.com/](http://nature.com/)

relativity100), such gravitational waves have never been observed directly — let alone used to study the cosmos. There are already Earth-based observatories hunting these waves, but a space-based one would search for waves at the opposite end of the spectrum (see *Nature* 525, 301–302; 2015). "It's like having a radio telescope as well as an optical one," says Karsten Danzmann, director of the Max Planck Institute for Gravitational Physics in Hanover, Germany, and co-principal investigator for the *Pathfinder* mission. "The part of the Universe you see is completely different."

The final space-based observatory will try to spot the stretching and compressing of space by bouncing laser beams between three masses floating in freefall, each separated from the others by some 5 million kilometres. Because the masses would be protected from all other

external forces, only a gravitational wave should disrupt the synchrony of their falling motion — a disturbance that would affect laser frequency.

The *LISA Pathfinder* (named after the Laser Interferometer Space Antenna, the concept behind the gravitational-wave observatory) is a smaller-scale test of this ultimate plan. With a pricetag of €400 million, it uses just two masses — each a 2-kilogram cube of gold and platinum — separated by a mere 38 centimetres, which allows them to fit inside the same spacecraft.

Unlike that of the observatory that it is designed to test-drive, this set-up is not sensitive enough to detect gravitational waves — instead, its purpose is to show that the masses can be completely isolated, and that any deviations in their relative motion can be measured with picometre accuracy. "We're missing out

the 5 million kilometres, but so what?" says Paul McNamara, the mission's project scientist. "Pretty much everything that could affect our ability to measure gravitational waves is here."

From the time of Pathfinder's launch from ESA's spaceport in Kourou, French Guiana, to the end of its subsequent eight-week journey, the masses will stay pinned to their housing deep inside the craft. But on arrival in orbit around a stable point between the Sun and Earth called Lagrange point 1, or L-1, about 1.5 million kilometres away, the cubes will be gently released to float within the spacecraft (see 'Precision lab in space').

Once in freefall, "the challenge is to isolate this little cube from everything around it, so the only thing it sees is space-time", says McNamara. Expected disturbances are pressure from solar radiation and stray magnetic fields; the equipment is so precise that it should detect even a force equal to the weight of a small bacterium on Earth.

As a high-precision laboratory in space, the LISA Pathfinder is unlike anything that ESA has done before, says Tim Sumner, an astrophysicist at Imperial College London who led the team that constructed one of the craft's protection mechanisms.

Another unusual element is that the major cargo — the cubes — will define the craft's trajectory, rather than vice versa. As they orbit around L-1 and fall in microgravity, Pathfinder will deploy microthrusters that are so gentle, it would take around 1,000 to lift a piece of paper on Earth. The thrusters will monitor the cubes' positions, ensuring that the craft hovers around the cubes without letting them touch its sides. Such a set-up required the teams who built the instruments and the engineers who made the craft to work together to an unprecedented degree, says Sumner.

These complexities go a long way towards explaining why the launch has taken so long to orchestrate, says Stefano Vitale, a physicist at the University of Trento in Italy, and a principal investigator for the Pathfinder mission; Pathfinder was approved by ESA in 2000 and originally intended for launch in 2006 (see *Nature* 469, 280; 2011). "Coarsely speaking, I think people underestimated the difficulty," says Vitale. "But that's why you have a Pathfinder."

The final step in the planned mission will test Pathfinder's limits by instructing onboard instruments to tweak the internal temperature

## PRECISION LAB IN SPACE

LISA Pathfinder aims to test whether an intricate experiment consisting of two metal cubes in freefall, isolated from all forces except gravity, can operate in space.

When Pathfinder launches, clamps pin the cubes — which are buried at the heart of the craft — tightly to their housing so that they don't jostle and damage either themselves or other instruments.

Two hours after launch, Pathfinder separates from the launcher and begins to make increasingly elongated ellipses.

Once the craft is stable, clamps release the cubes extremely gently; retractable devices position each one exactly at the centre of its housing at a speed of fewer than 5 micrometres per second. (See below).

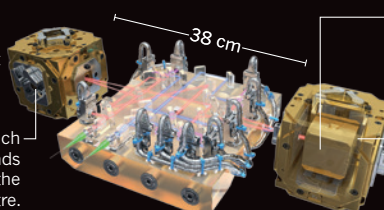
Around 55 days into the mission, craft arrives in orbit around L-1.

51 days after launch, Pathfinder separates from thrusters.

Nine days after launch, Pathfinder makes a final burn, propelling it towards its destination — the stable point L-1, 1.5 million kilometres from Earth.

At the heart of Pathfinder are two freefalling metal cubes, shielded from all forces except gravity by their housing.

The housing monitors each cube's position and commands the craft to move so that the cube is always at its centre.



Any disturbance to the relative motion of the cubes affects the frequency of the laser bouncing between them.

The cubes float in a vacuum, surrounded by instruments that mitigate stray forces.

and magnetic and electrostatic fields to see how such changes affect the cubes. "We want to learn everything we can about the physics of a free-floating body, and everything we learn will feed back into design of the future mission," says McNamara.

**"The challenge is to isolate this little cube from everything around it, so the only thing it sees is space-time."**

However, some opportunistic ESA scientists are already thinking about how Pathfinder's instruments could be used to inform other problems once its main mission, which could take up to a year, is complete. Measurement of the gravitational constant, known as Big G, for example, should fall naturally out of Pathfinder's data, Sumner says. Because the true value of Big G is disputed, a fresh measurement from

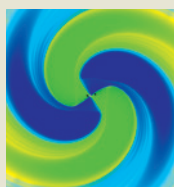
space would provide useful perspective.

To get a higher precision measurement, ESA may also consider extending the mission — although Sumner says that scientists would only request this after Pathfinder has proved itself, a few months in. He and his colleagues have also discussed using the craft's thrusters to send it to a spot known as a saddle point, where the gravitational pulls of Earth and the Sun cancel each other out. This could reveal how gravity behaves at its lowest level possible in the Solar System, with little extra cost. Few scientists doubt that Einstein's theories hold, says Sumner, but it would be interesting to do the test nonetheless.

Vitale, however, points out that it is important for researchers to stay focused on the mission's immediate goal. "Our main objective is to demonstrate freefall," he says, "and we don't want to be distracted from that." ■



### PHYSICS SPECIAL

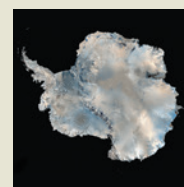


*Nature* celebrates 100 years of Einstein's general theory of relativity [nature.com/relativity100](http://nature.com/relativity100)

### MORE NEWS

- Safety upgrade found for gene-editing tweak [go.nature.com/rsnlqh](http://go.nature.com/rsnlqh)
- 'Supergene' determines sex strategy for wading birds [go.nature.com/uqaeyt](http://go.nature.com/uqaeyt)
- Engineered bat virus stirs up debate over risky research [go.nature.com/blqlcl](http://go.nature.com/blqlcl)

### LIVE CHAT



Glaciologist Matt Siegfried takes your questions on working in Antarctica [bit.ly/icediary](http://bit.ly/icediary)



## PHYSICS

# Satellites test general relativity

*Wayward craft find new use.*

BY ELIZABETH GIBNEY

Two satellites that were accidentally launched into the wrong orbit will be repurposed in the most stringent test yet of a prediction made by Albert Einstein's general theory of relativity — that clocks run more slowly the closer they are to heavy objects.

The satellites, operated by the European Space Agency (ESA), were mislaunched last year by a Russian Soyuz rocket that put them into elliptical, rather than circular, orbits. This left them unfit for their intended use as part of a European global-navigation system called Galileo. But the two craft have atomic clocks on board. According to general relativity, the clocks' ticking should slow down as the satellites move closer to Earth in their orbits, because the planet's gravity bends the fabric of space-time.

On 9 November, ESA announced that teams at Germany's Center of Applied Space Technology and Microgravity in Bremen and at the department of Time-Space Reference Systems at the Paris Observatory will track this acceleration and deceleration. By comparing the speed of the clocks' ticking with the crafts' known altitudes — pinpointed by lasers from monitoring stations on the ground — the researchers can test the accuracy of Einstein's theory.

In 1976, NASA launched an atomic clock aboard Gravity Probe A from Earth's surface, sending it 10,000 kilometres into space, to compare its ticking with that of an identical clock on the ground. But that probe stayed in the air for just shy of two hours. The Galileo satellites, by contrast, will conduct experiments for a year.

ESA expects the results to be four times more accurate than those of Gravity Probe A — enabling the agency to test whether theory agrees with reality to a precision of below 0.004%.

No one expects Einstein's theory, which was published almost 100 years ago (see [nature.com/relativity100](http://nature.com/relativity100)), to break down — it has passed every test thrown at it. A future ESA experiment called the Atomic Clock Ensemble in Space, or ACES, is scheduled to fly on the International Space Station in 2017. ACES will push Einstein's theory to even greater limits, testing it with a precision that could reach 0.0002%. ■



SUZANNE KREITER/THE BOSTON GLOBE/GETTY

Minority researchers in the United States consistently win NIH funding at lower rates than their peers.

## EQUALITY

# Racial bias haunts NIH grants

*Minorities are still less likely to get biomedical funding.*

BY ERIKA CHECK HAYDEN

Minority scientists are less likely than their peers to have biomedical research grants funded — and the disparity has barely changed in 30 years, according to data from the US National Institutes of Health (NIH). The numbers, requested by two California researchers to reignite discussion about diversity in the scientific workforce, show no consistent improvement, even though the proportion of minority grant reviewers has climbed.

Pulmonologist Esteban Burchard and epidemiologist Sam Oh of the University of California, San Francisco, shared the data with *Nature* after obtaining them from the NIH through a request under the Freedom of Information Act. The figures show that under-represented minorities have been awarded NIH grants at 78–90% the rate of white and mixed-race applicants every year from 1985 to 2013 (see 'Persistent gap').

Burchard and Oh had hypothesized that they might see an increase in funding for under-represented minorities after a 1994 NIH mandate that investigators must include women and minorities in clinical studies. They reasoned that the increased

emphasis on minority health would create a virtuous circle by boosting grants to minority researchers, who, they posit, would be more likely to focus on those groups and help to fulfil the mandate.

But there seems to have been no such increase. As a result, Burchard and Oh worry that a racial divide could develop between researchers and the people they study. Burchard notes, for example, that a lack of diversity among trial participants may have caused problems for two drug companies that produced an anticlotting drug that had reduced efficacy in East Asians and Pacific Islanders. The attorney-general of Hawaii filed a lawsuit against the companies last year for failing to disclose the issue. "It's a public-health problem," he says.

In a commentary in *PLoS Medicine* next month, Burchard and Oh will argue that scientific workforce diversity helps to ensure that research addresses issues relevant to all.

Other researchers say that airing the new data serves a purpose. "It raises the question in all of us as to the root causes of these disparities," says David Burgess, a cell biologist at Boston College in Chestnut Hill, Massachusetts, who is lead principal investigator of the National Research Mentoring Network,



SOURCE: E. BURCHARD/S. OH

a US\$19-million initiative announced by the NIH in October 2014 to improve mentorship of scientists from under-represented groups.

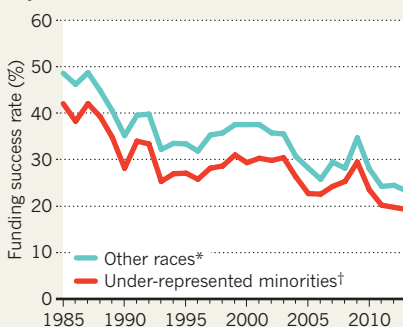
But the data offer no clues to such questions, counters Raynard Kington, president of Grinnell College in Iowa and former deputy director of the NIH. “It’s not surprising, not new, and doesn’t answer questions of how we can intervene to give every scientist the opportunity to contribute,” he says. In 2011, he co-authored a paper that found that black applicants for NIH funding were about two-thirds as likely as white people to receive grants during the years 2000–06, even accounting for factors such as publication record and training (D. K. Ginther *et al. Science* **333**, 1015–1019; 2011).

He and others point to evidence that funding can be influenced by personal bias. In February, researchers who analysed nearly 19,000 North American faculty hiring decisions in computer science, business and history reported that elite institutions predominantly hire people who earn their doctorates from the same or other elite schools. One-quarter of the 461 institutions surveyed had trained 71–86% of tenure-track faculty, depending on the discipline (A. Clauset *et al. Sci. Adv.* **1**, e1400005; 2015).

Such studies hold evidence of what biologist Margaret Werner-Washburne of the University of New Mexico in Albuquerque calls “positive bias” or in-group bias: the tendency for people

## PERSISTENT GAP

Since 1985, the chances of winning a US National Institutes of Health grant have fallen overall, but the success rate for under-represented minorities has stayed below that of other races.



\*White and mixed-race. †Pacific Islander, Native Hawaiian, African American, Native American and Asian.

to favour other people and institutions that they know either personally or by reputation.

“I think this happens a lot in the granting world,” she says. “Is it that when you’re on a panel and you have to rate 15 grants from the top people in the field who have really produced a lot, from major schools, you just want to root for them because their skill and potential is so apparent, versus someone who isn’t from that world?”

Cardiologist Hannah Valantine, who became

the NIH’s first chief officer for scientific-workforce diversity in 2014, says that the agency is focused on demonstrating the benefits of diversity and how to achieve it. She adds that in response to Kington’s 2011 paper, the NIH has allocated more than \$500 million to programmes to evaluate how to attract, mentor and retain minority researchers. The agency is also studying biases that might affect peer review, and is interested in gathering data on whether a diverse workforce improves science. Although diversity benefits businesses and individual scientific investigators, it has not been shown to broaden the scope of research, says Valantine.

“We can move forward with a premise that the diversity of scientists themselves is important,” she says. “But it behooves us as scientists to get the evidence that the diversity of scientists makes a difference to the output.” ■ [SEE EDITORIAL P.275](#)

## CLARIFICATION

The News story ‘Mega science prize split between more than 1,000 physicists’ (*Nature* **527**, 145; 2015) did not reflect Göran Hansson’s current role in the Nobel system. He is secretary-general of the Swedish Academy of Sciences, the body that awards the Nobel prizes in physics and chemistry.

# EXOPLANETS

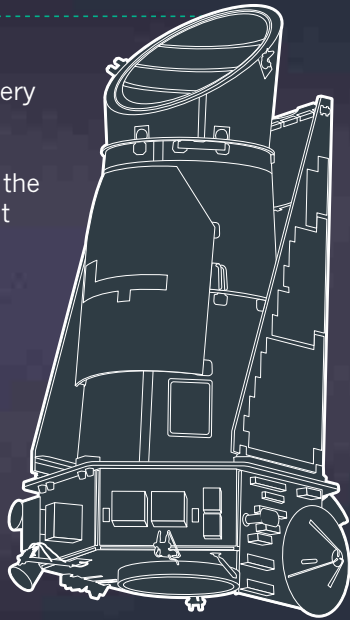
## THE NEXT 20 YEARS

Researchers have found nearly 2,000 worlds beyond our Solar System. Now they hope to understand them.

BY ALEXANDRA WITZE  
DESIGN BY JASIEK KRZYSZTOFIK

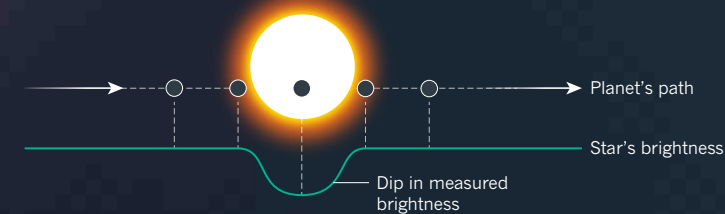
Twenty years ago this month, astronomers announced the discovery of 51 Pegasi b, the first confirmed planet orbiting a Sun-like star. The hellish gas giant orbits just beyond the searing heat of its parent star, and it opened astronomers' eyes to the astonishing range of alien worlds that exist throughout the Galaxy.

The tally of known extrasolar planets now stands at 1,978, with nearly 4,700 more candidates waiting to be confirmed. On 29 November, exoplanet researchers will gather in Hawaii to review these extreme solar systems — and map out a path for the next two decades.

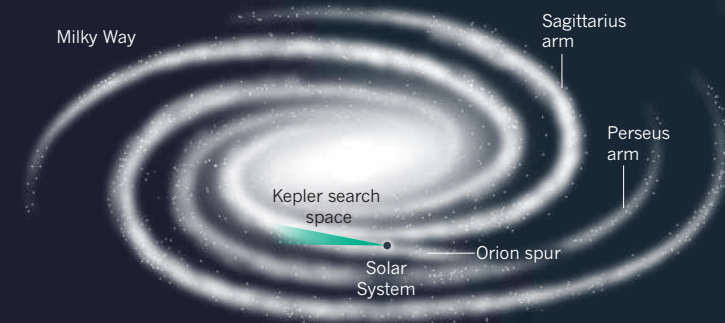


### The search so far

By far the greatest haul of exoplanets has come from NASA's Kepler spacecraft (pictured above), which for four years stared at a small patch of the night sky in search of stars that dim temporarily as planets cross their faces. The main Kepler mission ended in 2013, but planet hunting continues in a revamped 'K2' mission.

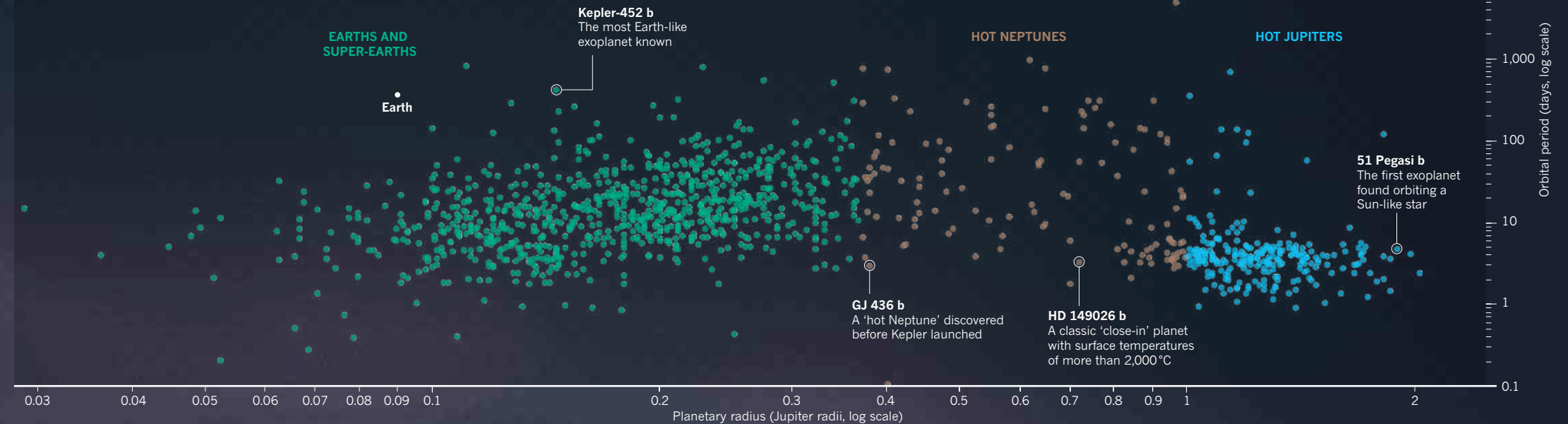


Kepler's field of view covers only about 1/400 of the night sky.



### THE WORLDS WE KNOW

Many of the exoplanets discovered to date are startlingly different from the worlds in the eight-planet architecture of our Solar System. They range from bloated gas balls close to their stars to ice worlds looping far beyond — and in between is a handful of Earth-like planets in the 'Goldilocks zone', where conditions are just right for life as scientists know it.



### THE NEXT FRONTIER

Astronomers now have to figure out what to do with this bonanza of planet discoveries. The research goals for the next two decades include gathering data on what the planets actually look like, from the clouds in their atmospheres to the conditions on their surfaces.

#### What's next?

##### GEMINI PLANET IMAGER

This mission is teasing out the heat of planets from that of their host stars, allowing direct measurements of characteristics such as mass, temperature and atmospheric composition.

##### NEXT-GENERATION TRANSIT SURVEY

An ongoing project to search for exoplanets in Southern Hemisphere skies.

##### TRANSITING EXOPLANET SURVEY SATELLITE

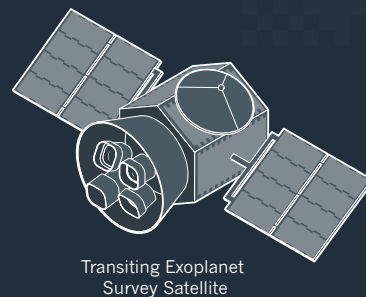
The spacecraft, set to launch in 2017, will search for rocky worlds around nearby bright stars. Astronomers can then follow up the finds using ground-based telescopes.

##### JAMES WEBB SPACE TELESCOPE

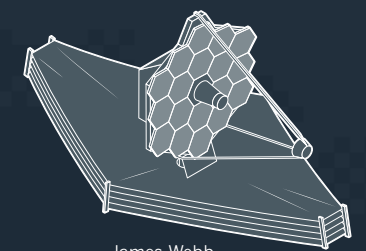
Targeted for a 2018 launch, the telescope will measure planetary atmospheres in infrared wavelengths to probe their chemical compositions.

##### PLATO

The space observatory, set to begin operating in 2024, will search for Earth-like worlds in the habitable zones of up to 1 million stars.



Transiting Exoplanet Survey Satellite



James Webb Space Telescope

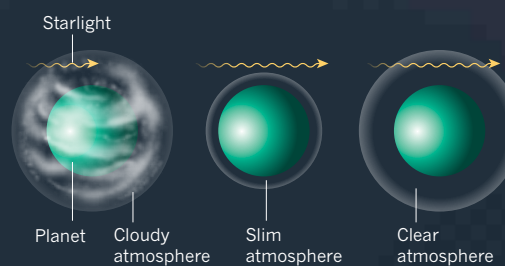
DATA FOR PLANETS SCATTERED BY EXOPLANETS; KEPLER PICTURE REDRAWN FROM NASA IMAGE

#### How many are there?

Untold numbers of exoplanets remain undiscovered, but astronomers are starting to get a better handle on the fraction of Earth-sized planets that might contain liquid water. The most common stars in the Galaxy are M dwarfs, which are smaller and cooler than the Sun; scientists estimate that there is up to one Earth-sized planet for every two M dwarfs. A fraction of those planets might be habitable.

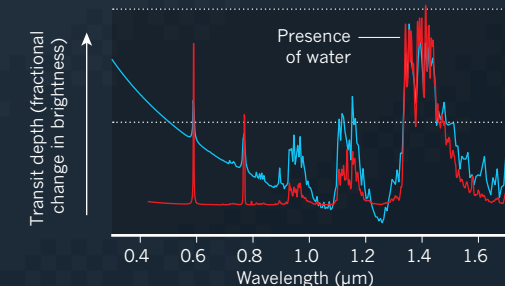
#### What do they look like?

The newest frontier is probing exoplanet atmospheres, looking at what changes as a planet slips on and off the face of its star (as seen from Earth).



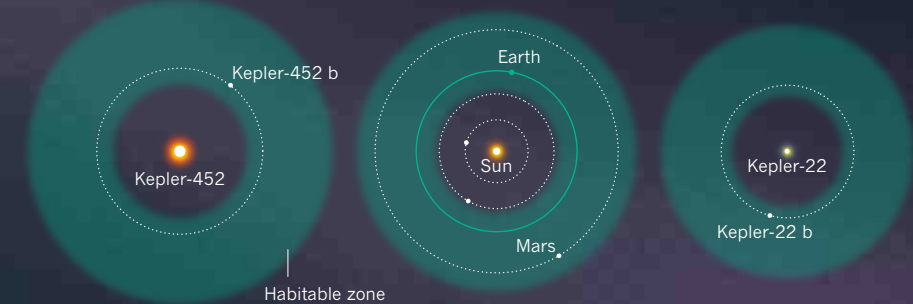
Chemical analyses of how the starlight is absorbed reveals compounds such as water in the cloudy skies of distant exoplanets.

— Model spectra with more haze  
— Model spectra with less haze



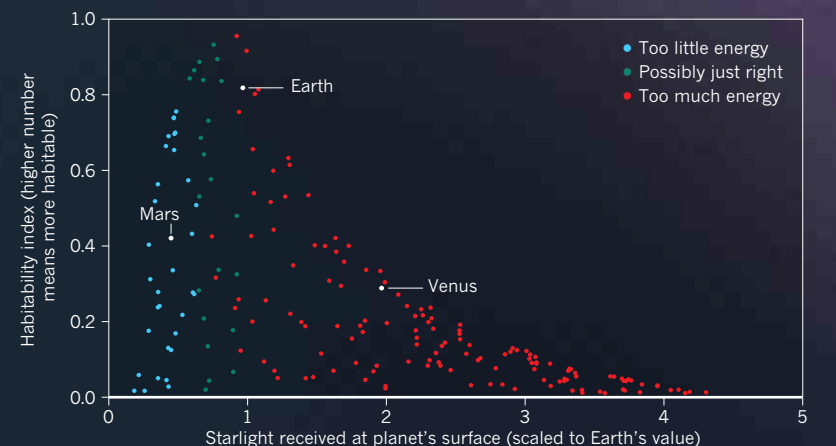
#### Are they habitable?

The most intriguing planets lie in the habitable zones of their stars, where temperatures allow liquid water to exist on the planet's surface. The placement and width of the habitable zone varies depending on how bright the host star is; the dimmer the star, the closer the planet must be to lie in the habitable zone.



#### So, is there life?

Maybe. Now the question is how to decide which of the potentially thousands of exoplanets to pursue further. Researchers recently devised a 'habitability index' that shows which planets are most likely to have liquid water on their surface. The index can be compared against other measures — such as the amount of starlight received by the planet — to explore which planets might be worth targeting first for searches for extraterrestrial life.





SPACE.

TIME.

## ENTANGLEMENT.

*Many physicists believe that entanglement is the essence of quantum weirdness — and some now suspect that it may also be the essence of space-time.*

BY RON COWEN

In early 2009, determined to make the most of his first sabbatical from teaching, Mark Van Raamsdonk decided to tackle one of the deepest mysteries in physics: the relationship between quantum mechanics and gravity. After a year of work and consultation with colleagues, he submitted a paper on the topic to the *Journal of High Energy Physics*.

In April 2010, the journal sent him a rejection — with a referee's report implying that Van Raamsdonk, a physicist at the University of British Columbia in Vancouver, was a crackpot.

His next submission, to *General Relativity and Gravitation*, fared little better: the referee's report was scathing, and the journal's editor asked for a complete rewrite.

But by then, Van Raamsdonk had entered a shorter version of the paper into a prestigious

annual essay contest run by the Gravity Research Foundation in Wellesley, Massachusetts. Not only did he win first prize, but he also got to savour a particularly satisfying irony: the honour included guaranteed publication in *General Relativity and Gravitation*. The journal published the shorter essay<sup>1</sup> in June 2010.

Still, the editors had good reason to be cautious. A successful unification of quantum mechanics and gravity has eluded physicists for nearly a century. Quantum mechanics governs the world of the small — the weird realm in which an atom or particle can be in many places at the same time, and can simultaneously spin both clockwise and anticlockwise. Gravity governs the Universe at large — from the fall of an apple to the motion of planets, stars and galaxies — and is described by Albert Einstein's

WARNER BROS. ENTERTAINMENT/  
PARAMOUNT PICTURES



Black holes such as the one depicted in *Interstellar* (2014) can be connected by wormholes.

Einstein loathed the idea of entanglement, and famously derided it as “spooky action at a distance”. But it is central to quantum theory. And Van Raamsdonk, drawing on work by like-minded physicists going back more than a decade, argued for the ultimate irony — that, despite Einstein’s objections, entanglement might be the basis of geometry, and thus of Einstein’s geometric theory of gravity. “Space-time,” he says, “is just a geometrical picture of how stuff in the quantum system is entangled.”

This idea is a long way from being proved, and is hardly a complete theory of quantum gravity. But independent studies have reached much the same conclusion, drawing intense interest from major theorists. A small industry of physicists is now working to expand the geometry–entanglement relationship, using all the modern tools developed for quantum computing and quantum information theory.

“I would not hesitate for a minute,” says physicist Bartłomiej Czech of Stanford University in California, “to call the connections between quantum theory and gravity that have emerged in the last ten years revolutionary.”

#### GRAVITY WITHOUT GRAVITY

Much of this work rests on a discovery<sup>2</sup> announced in 1997 by physicist Juan Maldacena, now at the Institute for Advanced Study in Princeton, New Jersey. Maldacena’s research had led him to consider the relationship between two seemingly different model universes. One is a cosmos similar to our own. Although it neither expands nor contracts, it has three dimensions, is filled with quantum particles and obeys Einstein’s equations of gravity. Known as anti-de Sitter space (AdS), it is commonly referred to as the bulk. The other model is also filled with elementary particles, but it has one dimension fewer and doesn’t recognize gravity. Commonly known as the boundary, it is a mathematically defined membrane that lies an infinite distance from any given point in the bulk, yet completely encloses it, much like the 2D surface of a balloon enclosing a 3D volume of air. The boundary particles obey the equations of a quantum system known as conformal field theory (CFT).

Maldacena discovered that the boundary and the bulk are completely equivalent. Like the 2D circuitry of a computer chip that encodes the 3D imagery of a computer game, the relatively simple, gravity-free equations that prevail on the boundary contain the same information and describe the same physics as the more complex equations that rule the bulk.

“It’s kind of a miraculous thing,” says Van Raamsdonk. Suddenly, he says, Maldacena’s duality gave physicists a way to think about quantum gravity in the bulk without thinking about gravity at all: they just had to look at the equivalent quantum state on the boundary.

And in the years since, so many have rushed to explore this idea that Maldacena’s paper is now one of the most highly cited articles in physics.

Among the enthusiasts was Van Raamsdonk, who started his sabbatical by pondering one of the central unsolved questions posed by Maldacena’s discovery: exactly how does a quantum field on the boundary produce gravity in the bulk? There had already been hints<sup>3</sup> that the answer might involve some sort of relation between geometry and entanglement. But it was unclear how significant these hints were: all the earlier work on this idea had dealt

*“I HAD UNDERSTOOD  
SOMETHING THAT  
NO ONE HAD  
UNDERSTOOD  
BEFORE.”*

with special cases, such as a bulk universe that contained a black hole. So Van Raamsdonk decided to settle the matter, and work out whether the relationship was true in general, or was just a mathematical oddity.

He first considered an empty bulk universe, which corresponded to a single quantum field on the boundary. This field, and the quantum relationships that tied various parts of it together, contained the only entanglement in the system. But now, Van Raamsdonk wondered, what would happen to the bulk universe if that boundary entanglement were removed?

He was able to answer that question using mathematical tools<sup>4</sup> introduced in 2006 by Shinsei Ryu, now at the University of Illinois at Urbana–Champaign, and Tadashi Takanagi, now at the Yukawa Institute for Theoretical Physics at Kyoto University in Japan. Their equations allowed him to model a slow and methodical reduction in the boundary field’s entanglement, and to watch the response in the bulk, where he saw space-time steadily elongating and pulling apart (see ‘The entanglement connection’). Ultimately, he found, reducing the entanglement to zero would break the space-time into disjointed chunks, like chewing gum stretched too far.

The geometry–entanglement relationship was general, Van Raamsdonk realized. Entanglement is the essential ingredient that knits space-time together into a smooth whole — not just in exotic cases with black holes, but always.

“I felt that I had understood something

general theory of relativity, announced 100 years ago this month. The theory holds that gravity is geometry: particles are deflected when they pass near a massive object not because they feel a force, said Einstein, but because space and time around the object are curved.

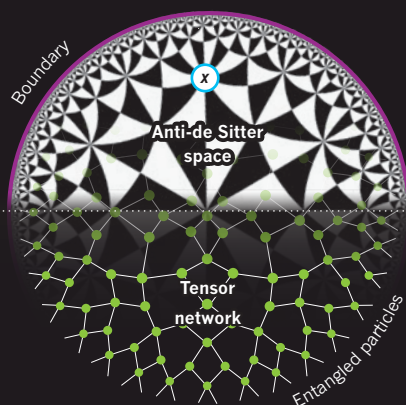
Both theories have been abundantly verified through experiment, yet the realities they describe seem utterly incompatible. And from the editors’ standpoint, Van Raamsdonk’s approach to resolving this incompatibility was strange. All that’s needed, he asserted, is ‘entanglement’: the phenomenon that many physicists believe to be the ultimate in quantum weirdness. Entanglement lets the measurement of one particle instantaneously determine the state of a partner particle, no matter how far away it may be — even on the other side of the Milky Way.

# THE ENTANGLEMENT CONNECTION

The ghostly quantum phenomenon of entanglement may be what knits space-time into a smooth whole.

In an infinite model universe known as anti-de Sitter space, the effects of gravity at any point  $x$  in the interior are mathematically equivalent to a quantum field theory on its boundary. This universe can be visualized in 2D by filling it with imaginary triangles. Although the triangles are identical, they look increasingly distorted as they approach the boundary.

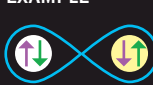
Physicists noticed that this pattern resembled diagrams called tensor networks, which were invented to show connections between quantum particles on a massive scale. These connections are known as quantum entanglement.



## What is quantum entanglement?

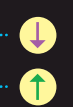
In 1935, Albert Einstein, Boris Podolsky and Nathan Rosen (EPR) pointed out that a connection can exist between widely separated quantum systems: a measurement of one will determine the state of the other.

### EXAMPLE



Entangled spins: if one particle is spinning up, the other spins down, and vice versa.

The particles are separated.

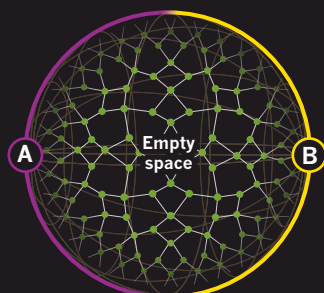


Observation of one particle instantaneously reveals the state of the other.

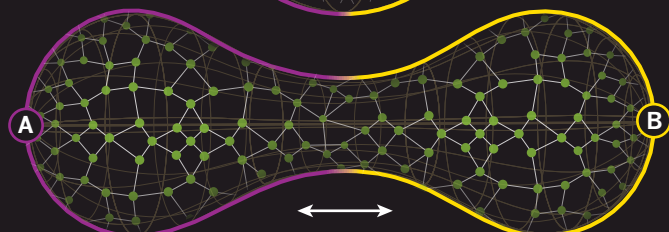
## DISENTANGLEMENT

The bulk-boundary correspondence implies that space on the inside is built from quantum entanglement around the outside.

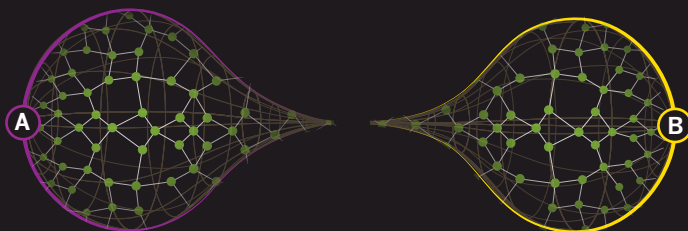
Even when the bulk universe is empty, the quantum fields in any two regions of the boundary (A and B) are heavily entangled with one another.



If the entanglement between these regions is reduced, the bulk universe starts pulling apart.



When the entanglement is reduced to zero, the bulk universe splits in two — showing that entanglement is necessary for space to exist.

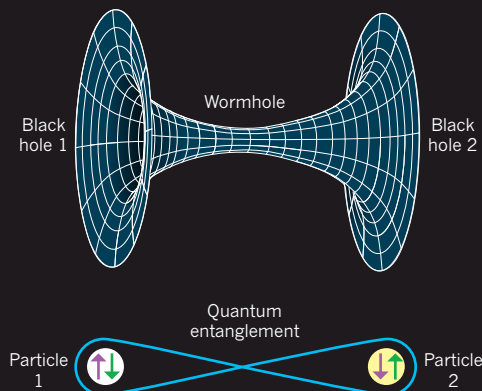


## ER = EPR

Also in 1935, Einstein and Rosen (ER) showed that widely separated black holes can be connected by a tunnel through space-time now often known as a wormhole.

Physicists suspect that the connection in a wormhole and the connection in quantum entanglement are the same thing, just on a vastly different scale.

Aside from their size there is no fundamental difference.



about a fundamental question that perhaps nobody had understood before,” he recalls: “Essentially, what is space-time?”

## ENTANGLEMENT AND EINSTEIN

Quantum entanglement as geometric glue — this was the essence of Van Raamsdonk’s rejected paper and winning essay, and an idea that has increasingly resonated among physicists. No one has yet found a rigorous proof, so the idea still ranks as a conjecture. But many independent lines of reasoning support it.

In 2013, for example Maldacena and Leonard Susskind of Stanford published<sup>5</sup> a related conjecture that they dubbed ER = EPR, in honour of two landmark papers from 1935. ER, by Einstein and American-Israeli physicist Nathan Rosen, introduced<sup>6</sup> what is now called a wormhole: a tunnel through space-time connecting two black holes. (No real particle could actually travel through such a wormhole, science-fiction films notwithstanding; that would require moving faster than light, which is impossible.) EPR, by Einstein, Rosen and American physicist Boris Podolsky, was the first paper to clearly articulate what is now called entanglement<sup>7</sup>.

Maldacena and Susskind’s conjecture was that these two concepts are related by more than a common publication date. If any two particles are connected by entanglement, the physicists suggested, then they are effectively joined by a wormhole. And vice versa: the connection that physicists call a wormhole is equivalent to entanglement. They are different ways of describing the same underlying reality.

No one has a clear idea of what this underlying reality is. But physicists are increasingly convinced that it must exist. Maldacena, Susskind and others have been testing the ER = EPR hypothesis to see if it is mathematically consistent with everything else that is known about entanglement and wormholes — and so far, the answer is yes.

## HIDDEN CONNECTIONS

Other lines of support for the geometry-entanglement relationship have come from condensed-matter physics and quantum information theory: fields in which entanglement already plays a central part. This has allowed researchers from these disciplines to attack quantum gravity with a whole array of fresh concepts and mathematical tools.

Tensor networks, for example, are a technique developed by condensed-matter physicists to track the quantum states of huge numbers of subatomic particles. Brian Swingle was using them in this way in 2007, when he was a graduate student at the Massachusetts Institute of Technology (MIT) in Cambridge, calculating how groups of electrons interact in a solid material. He found that the most useful network for this purpose started by linking adjacent pairs of electrons, which are most likely to interact with each other, then linking larger and larger groups in a pattern that resembled the hierarchy



of a family tree. But then, during a course in quantum field theory, Swingle learned about Maldacena's bulk–boundary correspondence and noticed an intriguing pattern: the mapping between the bulk and the boundary showed exactly the same tree-like network.

Swingle wondered whether this resemblance might be more than just coincidence. And in 2012, he published<sup>8</sup> calculations showing that it was: he had independently reached much the same conclusion as Van Raamsdonk, thereby adding strong support to the geometry–entanglement idea. “You can think of space as being built from entanglement in this very precise way using the tensors,” says Swingle, who is now at Stanford and has seen tensor networks become a frequently used tool to explore the geometry–entanglement correspondence.

Another prime example of cross-fertilization is the theory of quantum error-correcting codes, which physicists invented to aid the construction of quantum computers. These machines encode information not in bits but in ‘qubits’: quantum states, such as the up or down spin of an electron, that can take on values of 1 and 0 simultaneously. In principle, when the qubits interact and become entangled in the right way, such a device could perform calculations that an ordinary computer could not finish in the lifetime of the Universe. But in practice, the process can be incredibly fragile: the slightest disturbance from the outside world will disrupt the qubits’ delicate entanglement and destroy any possibility of quantum computation.

That need inspired quantum error-correcting codes, numerical strategies that repair corrupted correlations between the qubits and make the computation more robust. One hallmark of these codes is that they are always ‘non-local’: the information needed to restore any given qubit has to be spread out over a wide region of space. Otherwise, damage in a single spot could destroy any hope of recovery. And that non-locality, in turn, accounts for the fascination that many quantum information theorists feel when they first encounter Maldacena's bulk–boundary correspondence: it shows a very similar kind of non-locality. The information that corresponds to a small region of the bulk is spread over a vast region of the boundary.

“Anyone could look at AdS–CFT and say that it's sort of vaguely analogous to a quantum error-correcting code,” says Scott Aaronson, a computer scientist at MIT. But in work published in June<sup>9</sup>, physicists led by Daniel Harlow at Harvard University in Cambridge and John Preskill of the California Institute of Technology in Pasadena argue for something stronger: that the Maldacena duality is itself a quantum error-correcting code. They have demonstrated that this is mathematically correct in a simple model, and are now trying to show that the assertion holds more generally.

“People have been

saying for years that entanglement is somehow important for the emergence of the bulk,” says Harlow. “But for the first time, I think we are really getting a glimpse of how and why.”

## BEYOND ENTANGLEMENT

That prospect seems to be enticing for the Simons Foundation, a philanthropic organization in New York City that announced in August that it would provide US\$2.5 million per year for at least 4 years to help researchers to move forward on the gravity–quantum information connection. “Information theory provides a powerful way to structure our thinking about fundamental physics,” says Patrick Hayden, the Stanford physicist who is directing the programme. He adds that the Simons sponsorship

# “YOU CAN THINK OF SPACE AS BEING BUILT FROM ENTANGLEMENT.”

will support 16 main researchers at 14 institutions worldwide, along with students, postdocs and a series of workshops and schools. Ultimately, one major goal is to build up a comprehensive dictionary for translating geometric concepts into quantum language, and vice versa. This will hopefully help physicists to find their way to the complete theory of quantum gravity.

Still, researchers face several challenges. One is that the bulk–boundary correspondence does not apply in our Universe, which is neither static nor bounded; it is expanding and apparently infinite. Most researchers in the field do think that calculations using Maldacena's correspondence are telling them something true about the real Universe, but there is little agreement as yet on exactly how to translate results from one regime to the other.

Another challenge is that the standard definition of entanglement refers to particles only at a given moment. A complete theory of quantum gravity will have to add time to that picture. “Entanglement is a big piece of the story, but it's not the whole story,” says Susskind.

He thinks physicists may have to embrace another concept from quantum information theory: computational complexity, the number of logical steps, or operations, needed to construct the quantum state of a system. A system with low complexity is analogous to a quantum computer with almost all the qubits on zero: it is easy to define and to build. One with high complexity is analogous to a set of qubits encoding a

number that would take aeons to compute.

Susskind began to think about computational complexity about a decade ago, when he noticed that a solution to Einstein's equations of general relativity allowed a wormhole in AdS space to get longer and longer as time went on. What did that correspond to on the boundary, he wondered? What was changing there? Susskind knew that it couldn't be entanglement, because the correlations that produce entanglement between different particles on the boundary reach their maximum in less than a second<sup>10</sup>. In an article last year<sup>11</sup>, however, he and Douglas Stanford, now at the Institute for Advanced Study, showed that as time progressed, the quantum state on the boundary would vary in exactly the way expected from computational complexity.

“It appears more and more that the growth of the interior of a black hole is exactly the growth of computational complexity,” says Susskind. If quantum entanglement knits together pieces of space, he says, then computational complexity may drive the growth of space — and thus bring in the elusive element of time. One potential consequence, which he is just beginning to explore, could be a link between the growth of computational complexity and the expansion of the Universe. Another is that, because the insides of black holes are the very regions where quantum gravity is thought to dominate, computational complexity may have a key role in a complete theory of quantum gravity.

Despite the remaining challenges, there is a sense among the practitioners of this field that they have begun to glimpse something real and very important. “I didn't know what space was made of before,” says Swingle. “It wasn't clear that question even had meaning.” But now, he says, it is becoming increasingly apparent that the question does make sense. “And the answer is something that we understand,” says Swingle. “It's made of entanglement.”

As for Van Raamsdonk, he has written some 20 papers on quantum entanglement since 2009. All of them, he says, have been accepted for publication. ■

**Ron Cowen** is a freelance writer based in Silver Spring, Maryland.

1. Van Raamsdonk, M. *Gen. Relativ. Grav.* **42**, 2323–2329 (2010).
2. Maldacena, J. M. *Adv. Theor. Math. Phys.* **2**, 231–252 (1998).
3. Maldacena, J. M. *J. High Energy Phys.* **2003**, 021 (2003).
4. Ryu, S. & Takayanagi, T. *Phys. Rev. Lett.* **96**, 181602 (2006).
5. Maldacena, J. & Susskind, L. *Fortschr. Phys.* **61**, 781–811 (2013).
6. Einstein, A. & Rosen, N. *Phys. Rev.* **48**, 73–77 (1935).
7. Einstein, A., Podolsky, B. & Rosen, N. *Phys. Rev.* **47**, 777–780 (1935).
8. Swingle, B. *Phys. Rev. D* **86**, 065007 (2012).
9. Pastawski, F. et al. *J. High Energy Phys.* **2015**, 149 (2015).
10. Susskind, L. Preprint at <http://arxiv.org/abs/1411.0690> (2014).
11. Stanford, D. & Susskind, L. *Phys. Rev. D* **90**, 126007 (2014).

**➔ NATURE.COM**  
For more on general relativity at 100.  
[nature.com/relativity100](http://nature.com/relativity100)

# COMMENT

**HISTORY** Friends and rivals fed into Einstein's general theory of relativity **p.298**



**MATHEMATICS** The playful and prescient logic of *Alice in Wonderland* **p.302**

**EMISSIONS** Burning peat to plant oil palms in Indonesia has to stop **p.305**

**OBITUARY** Richard Heck, palladium-catalysis pioneer, remembered **p.306**

XINHUA/XINHUA PRESS/CORBIS



The central route of China's South-to-North Water Diversion project runs through Jiaozuo in Henan province.

## Transfer project cannot meet China's water needs

Better local water management is the way to keep pace with escalating demands, not pumping water across the country, warn **Jon Barnett** and colleagues.

**A**lmost one year ago, Beijing began to receive water channelled by the South-to-North Water Diversion (SNWD) project. The biggest inter-basin transfer scheme in the world, the SNWD project has the capacity to deliver 25 billion cubic metres of fresh water per year from the Yangtze River in China's south to the drier north by two routes — each of which covers a distance of more than 1,000 kilometres. The project connects four major river basins, three megacities, six provinces and hundreds of millions of water users and polluters.

Its success is already in question. Reservoir and canal construction costs have reportedly reached US\$80 billion, and more than 300,000 people have been displaced<sup>1</sup>. Pollution and environmental fallout, as well as high maintenance costs and water prices, make the project unsustainable both ecologically and socially. And the transfer of water does not address the underlying causes of water shortages in the north, namely pollution and inefficient agricultural, industrial and urban use — the effects of which we have been studying over the past decade.

North China could be self-sufficient in water without the transfer of water from the south. But the necessary steps — among them, improving local pollution monitoring and building better irrigation infrastructure — are inadequately implemented.

Increasing supply is viewed as the main solution to water scarcity because of the conflicting roles of the Chinese government as both entrepreneur and regulator. Incentives for economic growth in China still outweigh incentives for pollution control and limits on water extraction, despite ever stricter ►



► environmental laws. Many industries, such as the country's huge hydropower sector, profit from expensive solutions to boosting water supplies.

China's water system needs an overhaul. Institutional reforms must divorce profit motives from regulatory functions; data and decisions must be disclosed to the public; and the influence of the hydropower sector on water-resource management needs to be restricted. The volume of water being diverted along existing routes of the SNWD project must be reduced and extensions to the project must be shelved.

Better local management of resources is the only way to bring secure and sustainable water to all parts of China.

## BIG WATER

China's history of grand water-engineering projects is almost as old as the nation itself, and is inextricably knit with the country's politics, development and self-image. The first dam was built in around 600 BC at Anfeng Tang in eastern China. It created a still existing reservoir 100 kilometres in circumference that could irrigate an area of 24,000 square kilometres. Ever since, most of China's water-management systems have been created and run by the state.

The SNWD project transports water in two ways (see 'South-to-north water transfer'). Its eastern route has the capacity to supply up to 14.8 billion cubic metres of water per year to the provinces of Jiangsu, Anhui, Shandong and Hebei, and to the city of Tianjin. The water travels through a system of pumps, rivers, lakes, reservoirs and canals that includes the Grand Canal, which was built around 500 BC. Its central route will provide up to 9.5 billion cubic metres of water per year, including one-third of Beijing's water, from the Danjiangkou reservoir on the Han River (a tributary of the Yangtze). During the construction of this route, the water level of the reservoir was raised by 13 metres, which resulted in the resettlement of 180,000 people from Shiyan city and 160,000 from Nanyang city.

A third, western, route is planned that would divert up to 20 billion cubic metres of water from three tributaries of the upper Yangtze through tunnels to the upper reaches of the Yellow River. Its path is under debate and there has been no commitment to commencing its construction nor any indication of when a decision might be made<sup>1</sup>. In our view, the scope for improving water management makes this extra route unnecessary.

Without question, northern China, which includes the Hai, Huai and Yellow river basins, is short of water. The region's annual per capita water availability is only around half of the international threshold for water stress<sup>2</sup>. Water scarcity is most acute in the Hai basin, where Beijing is located. Farms and cities

## SOUTH-TO-NORTH WATER TRANSFER

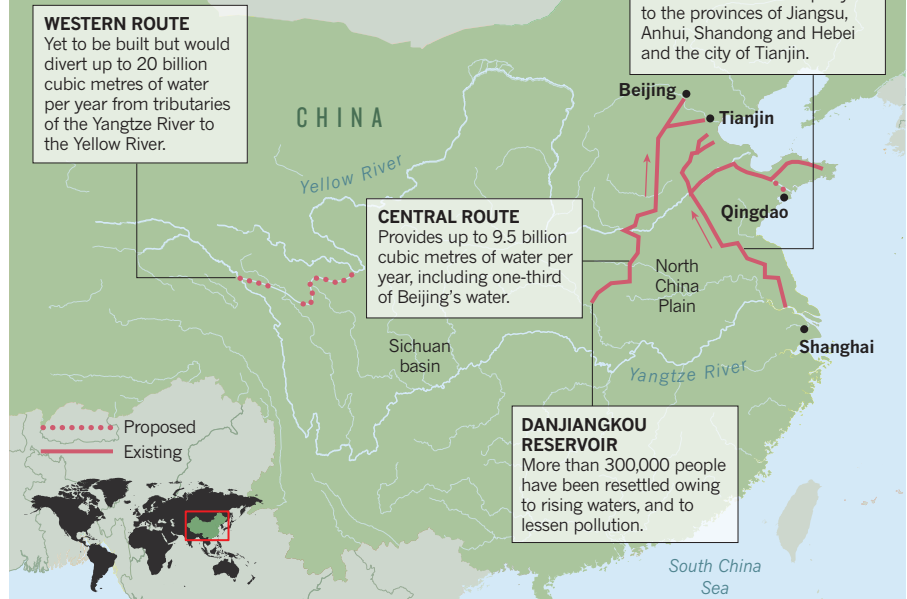
Mainland China has the capacity to pump 25 billion cubic metres of fresh water per year over a distance of more than 1,000 kilometres — from the Yangtze River in China's south to Beijing and other cities in the drier north.

**WESTERN ROUTE**  
Yet to be built but would divert up to 20 billion cubic metres of water per year from tributaries of the Yangtze River to the Yellow River.

**CENTRAL ROUTE**  
Provides up to 9.5 billion cubic metres of water per year, including one-third of Beijing's water.

**EASTERN ROUTE**  
Supplies up to 14.8 billion cubic metres of water per year to the provinces of Jiangsu, Anhui, Shandong and Hebei and the city of Tianjin.

**DANJIANGKOU RESERVOIR**  
More than 300,000 people have been resettled owing to rising waters, and to lessen pollution.



have increasingly drawn on groundwater such that 50% of aquifers in the North China Plain are now below sea level. This scarcity is compounded by poor water quality in up to 60% of water in the rivers of the north, which further reduces the supply of clean water for drinking and domestic use<sup>2</sup>.

The problem is more a scarcity of management than of natural water<sup>3</sup>. Inefficient agricultural production consumes about 75% of the region's water and is growing rapidly. In areas where cereal crops are flood irrigated, losses can exceed 50%. In addition, the lack of storage systems along the Yellow River means that farmers must use water when it is made available — not necessarily when they need it<sup>3</sup>.

The idea of water scarcity in the north is perpetuated by China's government for several reasons. It justifies taking water from the south to achieve President Xi Jinping's ambitions for a mega-economic region that encompasses Beijing, Tianjin and Hebei. And it serves the interests of those in the business of supplying water, including China's huge, state-owned water-engineering firms.

But the SNWD project does not ensure a reliable supply to the north. Pollution is a pervasive risk. In response to complaints about quality from provinces receiving water, the National Development and Reform Commission ordered changes in land use across the Danjiangkou reservoir catchment area to reduce urban and agricultural run-off. Development has been prohibited in some areas, and in others communities have been resettled. The use of pesticides and fertilizers has been limited and industry is subjected to

stricter pollution controls. In 2015, Danjiangkou reservoir won a national award for water quality — at the cost of the impoverishment of the hundreds of thousands of people who were forced to move.

And there are other costs. Wang Mengshu, a civil engineer at Beijing Jiaotong University, has suggested that the expense of maintaining the SNWD conduits was vastly underestimated. The price of transferred water will be too expensive for farmers, who will therefore continue to exploit groundwater<sup>4</sup>.

The SNWD project also poses risks in source areas. Claims of abundant water in the Yangtze hide the fact that shortages do occur. In the past decade, there have been two severe droughts in the Yangtze basin. And periods of water scarcity are more likely in the future because of an increase in the number of withdrawals and dams, as well as the effects of urbanization and climate change. The timing of water transfers is therefore important: should extractions from the Yangtze occur at times of low flow, saline waters from its estuary could be drawn in. Nearby Shanghai's population of 24 million would then face critical water shortages until discharge levels rose again<sup>5</sup>.

Governance of the SNWD project remains unresolved. Both the state-owned HydroChina Corporation and the central government's SNWD project construction committee seek to control the flow of the project's waters, even though this is a core responsibility of the Chinese Ministry of Water Resources. Corporatization of the state is reaching into the management of water,

SOURCE: OFFICE OF THE SNWD PROJECT COMMISSION OF THE STATE COUNCIL/J.B.



Hundreds of thousands of residents were relocated during construction of the project.

TAO DEBIN/XINHUA PRESS/CORBIS

creating tension between motives and profit, uncertainty about roles and responsibilities and impediments to coordinated management of the nation's water courses.

### BEYOND PIPES

In cities, rainwater harvesting and wastewater recycling can meet much of the demand. According to Qiu Baoxing, a former vice-minister of the Chinese Ministry of Housing and Urban–Rural Development, the SNWD project could have been avoided if one-third of buildings in Beijing collected rainwater. Increased investment in treatment systems, efficient irrigation and the monitoring and enforcement of pollution levels can also improve the supply of usable water<sup>2,6,7</sup>. Lower-quality water could be used for urban landscaping and industry, and some water-intensive activities could be moved to the south.

Such solutions require coordination with local governments, which are driven by growth and profit. When bureaucrats behave like businessmen and state-owned enterprises operate like private corporations, even strong environmental laws have little effect<sup>6</sup>. By contrast, the SNWD project is easy to administer, politically feasible and drives growth.

There are signs of change. Since 2006, environmental targets have been included in performance criteria for local leaders<sup>6</sup>. And there have been experiments in increasing disclosure to the public of data on the environmental performance of firms<sup>7</sup>. Both measures have made local governments and businesses

more accountable for environmental standards. They have led to lower levels of pollution and encouraged investment in cleaner technologies and the closure of inefficient plants. But the enforcement of standards and laws varies from region to region — the economic imperative still dominates in less-developed regions — and data can be falsified<sup>6</sup>.

The Chinese government's authority rests on maintaining social stability and economic growth. The government must therefore respond to challenges such as corruption, public-health issues and inequality<sup>8</sup>. Given improved living standards, greater levels of education and the proliferation of social media, high levels of pollution can no longer be ignored. Both the central and local governments in China must be seen to be controlling pollution, which can lead to secrecy and misinformation. In our experience, detailed data on the flow of water and pollution levels in major rivers can be difficult to obtain and must often be paid for.

### SMALL STEPS

As is already done for air pollution, the central and local Chinese governments should disclose information on water to demonstrate the responsible management of resources to the public. Providing accessible information about the allocation of water rights — as well as the allocation of water itself to

***"The problem is more a scarcity of management than of natural water."***

provinces, irrigation districts and farmers — would increase public trust in the system and improve the accountability of water managers, local government and firms<sup>6</sup>.

Local environmental-protection bureaus should be given the autonomy and resources to collect and analyse monitoring data independently and to enforce pollution standards. Exporters that rely on foreign investment must increasingly comply with standards and regulations as their parent companies and consumers demand proof of environmental responsibility. Industrial water users should consider cleaner production as a path to savings, new markets and improved competitiveness<sup>9</sup>.

In agriculture, losses can be reduced by lining irrigation canals with concrete. Water should be supplied only at times when irrigation is necessary<sup>3</sup>. The rotation of wheat with higher-value crops that take less water to grow, such as peanuts, will also improve the efficiency of water use<sup>10</sup>.

Investment in new technologies is needed, including systems to separate urban water according to quality, recycle waste water, encourage water conservation and improve the harvesting of rainwater. This would require performance targets to be set for local managers, as well as investment in and incentives for building smaller-scale water infrastructure. Campaigns to increase public awareness of water issues should also be implemented.

Constraining the influence of the hydropower sector on water-resource management will help to shift public investment towards these smaller-scale technologies. The sector is already expanding into overseas markets to compensate for reduced domestic demand in the wake of disquiet about water pollution.

As its limitations become clear, the SNWD project might well mark the nadir of big-engineering solutions to China's water problems. ■

**Jon Barnett** is professor and Australian Research Council Future Fellow, **Sarah Rogers** is a research fellow, and **Michael Webber, Brian Finlayson and Mark Wang** are professors, in the School of Geography, University of Melbourne, Victoria, Australia. email: [jbarn@unimelb.edu.au](mailto:jbarn@unimelb.edu.au)

1. Crow-Miller, B. *Water Altern.* **8**, 173–192 (2015).
2. Jiang, Y. *Environ. Sci. Policy* **54**, 106–125 (2015).
3. Webber, M., Barnett, J., Finlayson, B. & Wang, M. *Global Environ. Chang.* **18**, 617–625 (2008).
4. Webber, M. *Making Capitalism in Rural China* (Edward Elgar, 2012).
5. Chen, D. et al. *Appl. Geogr.* **45**, 303–310 (2013).
6. Wang, A. L. *Harvard Environ. Law Rev.* **37**, 365–440 (2013).
7. Wang, H. et al. *J. Environ. Mgmt* **71**, 123–133 (2004).
8. Heberer, T. & Schubert, G. (eds) *Regime Legitimacy in Contemporary China: Institutional Change and Stability* (Routledge, 2009).
9. Yee, W.-H., Lo, C. W.-H. & Tang, S.-Y. *China Quart.* **213**, 101–129 (2013).
10. Yang, X. et al. *PLoS ONE* **10**, e0115269 (2015).





GROSSMANN, EINSTEIN: ETH-BIBLIOTHEK ZÜRICH/BIL DARCHIV;  
BESSO: BESSO FAMILY/AP EMILIO SEGRE VISUAL ARCHIVES

Marcel Grossmann (left) and Michele Besso (right), university friends of Albert Einstein (centre), both made important contributions to general relativity.

# Einstein was no lone genius

Lesser-known and junior colleagues helped the great physicist to piece together his general theory of relativity, explain **Michel Janssen** and **Jürgen Renn**.

A century ago, in November 1915, Albert Einstein published his general theory of relativity in four short papers in the proceedings of the Prussian Academy of Sciences in Berlin<sup>1</sup>. The landmark theory is often presented as the work of a lone genius. In fact, the physicist received a great deal of help from friends and colleagues, most of whom never rose to prominence and have been forgotten<sup>2–5</sup>. (For full reference details of all Einstein texts mentioned in this piece, see Supplementary Information; [go.nature.com/ufcgp9](http://go.nature.com/ufcgp9).)

Here we tell the story of how their insights were woven into the final version of the theory. Two friends from Einstein's student days — Marcel Grossmann and Michele Besso — were particularly important. Grossmann was a gifted mathematician and organized student who helped the more visionary and fanciful Einstein at crucial moments. Besso was an engineer, imaginative and somewhat disorganized, and a caring and lifelong friend to Einstein. A cast of others contributed too.

Einstein met Grossmann and Besso at the Swiss Federal Polytechnical School in Zurich<sup>6</sup> — later renamed the Swiss Federal Institute of Technology (Eidgenössische

Technische Hochschule; ETH) — where, between 1896 and 1900, he studied to become a school teacher in physics and mathematics. Einstein also met his future wife at the ETH, classmate Mileva Marić. Legend has it that Einstein often skipped class and relied on Grossmann's notes to pass exams.

Grossmann's father helped Einstein to secure a position at the patent office in Berne in 1902, where Besso joined him two years later. Discussions between Besso and Einstein earned the former the sole acknowledgment in the most famous of Einstein's 1905 papers, the one introducing the special theory of relativity. As well as publishing the papers that made 1905 his *annus mirabilis*, Einstein completed his dissertation that year to earn a PhD in physics from the University of Zurich.

In 1907, while still at the patent office, he started to think about extending the principle of relativity from uniform to arbitrary motion through a new theory of gravity. Presciently, Einstein wrote to his friend Conrad Habicht — whom he knew from a reading group in Berne mockingly called the Olympia Academy by its three members — saying that he hoped that this new theory

would account for a discrepancy of about 43" (seconds of arc) per century between Newtonian predictions and observations of the motion of Mercury's perihelion, the point of its orbit closest to the Sun.

Einstein started to work in earnest on this new theory only after he left the patent office in 1909, to take up professorships first at the University of Zurich and two years later at the Charles University in Prague. He realized that gravity must be incorporated into the structure of space-time, such that a particle subject to no other force would follow the straightest possible trajectory through a curved space-time.

In 1912, Einstein returned to Zurich and was reunited with Grossmann at the ETH. The pair joined forces to generate a fully fledged theory. The relevant mathematics was Gauss's theory of curved surfaces, which Einstein probably learned from Grossmann's notes. As we know from recollected conversations, Einstein told Grossmann<sup>7</sup>: "You must help me, or else I'll go crazy."

Their collaboration, recorded in Einstein's 'Zurich notebook', resulted in a joint paper published in June 1913, known as the Entwurf ('outline') paper. The main advance between this 1913 Entwurf

theory and the general relativity theory of November 1915 are the field equations, which determine how matter curves space-time. The final field equations are 'generally covariant': they retain their form no matter what system of coordinates is chosen to express them. The covariance of the Entwurf field equations, by contrast, was severely limited.

## TWO THEORIES

In May 1913, as he and Grossmann put the finishing touches to their Entwurf paper, Einstein was asked to lecture at the annual meeting of the Society of German Natural Scientists and Physicians to be held that September in Vienna, an invitation that reflects the high esteem in which the 34-year-old was held by his peers.

In July 1913, Max Planck and Walther Nernst, two leading physicists from Berlin, came to Zurich to offer Einstein a well-paid and teaching-free position at the Prussian Academy of Sciences in Berlin, which he swiftly accepted and took up in March 1914. Gravity was not a pressing problem for Planck and Nernst; they were mainly interested in what Einstein could do for quantum physics.

Several new theories had been proposed in which gravity, like electromagnetism, was represented by a field in the flat space-time of special relativity. A particularly promising one came from the young Finnish physicist Gunnar Nordström. In his Vienna lecture, Einstein compared his own Entwurf theory to Nordström's theory. Einstein worked on both theories between May and late August 1913, when he submitted the text of his lecture for publication in the proceedings of the 1913 Vienna meeting.

In the summer of 1913, Nordström visited Einstein in Zurich. Einstein convinced him that the source of the gravitational field in both their theories should be constructed out of the 'energy-momentum tensor': in pre-relativistic theories, the density and the flow of energy and momentum were represented by separate quantities; in relativity theory, they are combined into one quantity with ten different components.

This energy-momentum tensor made its first appearance in 1907–8 in the special-relativistic reformulation of the theory of electrodynamics of James Clerk Maxwell and Hendrik Antoon Lorentz by Hermann Minkowski. It soon became clear that an energy-momentum tensor could be defined for physical systems other than electromagnetic fields. The tensor took centre stage in the new relativistic mechanics presented in the first textbook on special relativity, *Das Relativitätssprinzip*, written by Max Laue in 1911. In 1912, a young Viennese physicist, Friedrich Kottler, generalized Laue's formalism from flat to curved space-time. Einstein and Grossmann relied on this generalization

in their formulation of the Entwurf theory. During his Vienna lecture, Einstein called for Kottler to stand up and be recognized for this work<sup>8</sup>.

Einstein also worked with Besso that summer to investigate whether the Entwurf theory could account for the missing 43" per century for Mercury's perihelion. Unfortunately, they found that it could only explain 18". Nordström's theory, Besso checked later, gave 7" in the wrong direction. These calculations are preserved in the 'Einstein-Besso manuscript' of 1913.

Besso contributed significantly to the calculations and raised interesting questions. He wondered, for instance, whether the Entwurf field equations have an unambiguous solution that uniquely determines the gravitational field of the Sun. Historical analysis of extant manuscripts suggests that this query gave Einstein the idea for an argument that reconciled him with the restricted covariance of the Entwurf equations. This 'hole argument' seemed to show that generally covariant field equations cannot uniquely determine the gravitational field and are therefore inadmissible<sup>9</sup>.

Einstein and Besso also checked whether the Entwurf equations hold in a rotating coordinate system. In that case the inertial forces of rotation, such as the centrifugal force we experience on a merry-go-round, can be interpreted as gravitational forces. The theory seemed to pass this test. In August 1913, however, Besso warned him that it did not. Einstein did not heed the warning, which would come back to haunt him.

In his lecture in Vienna in September 1913, Einstein concluded his comparison of the two theories with a call for experiment to decide. The Entwurf theory predicts that gravity bends light, whereas Nordström's does not. It would take another five years to find out. Erwin Finlay Freundlich, a junior astronomer in Berlin with whom Einstein had been in touch since his days in Prague, travelled to Crimea for the solar eclipse of August 1914 to determine whether gravity bends light but was interned by the Russians just as the First World War broke out. Finally, in 1919, English astronomer Arthur Eddington confirmed Einstein's prediction of light bending by observing the deflection of distant stars seen close to the Sun's edge during another eclipse, making Einstein a household name<sup>10</sup>.

Back in Zurich, after the Vienna lecture, Einstein teamed up with another young physicist, Adriaan Fokker, a student of Lorentz, to reformulate the Nordström

theory using the same kind of mathematics that he and Grossmann had used to formulate the Entwurf theory. Einstein and Fokker showed that in both theories the gravitational field can be incorporated into the structure of a curved space-time. This work also gave Einstein a clearer picture of the structure of the Entwurf theory, which helped him and Grossmann in a second joint paper on the theory. By the time it was published in May 1914, Einstein had left for Berlin.

## THE BREAKTHROUGH

Turmoil erupted soon after the move. Einstein's marriage fell apart and Mileva moved back to Zurich with their two young sons. Albert renewed the affair he had started and broken off two years before with his cousin Elsa Löwenthal (née Einstein). The First World War began. Berlin's scientific elite showed no interest in the Entwurf theory, although renowned colleagues elsewhere did, such as Lorentz and Paul Ehrenfest in Leiden, the Netherlands. Einstein soldiered on.

By the end of 1914, his confidence had grown enough to write a long exposition of the theory. But in the summer of 1915, after a series of his lectures in Göttingen had piqued the interest of the great mathematician David Hilbert, Einstein started to have serious doubts. He discovered to his dismay that the Entwurf theory does not make rotational motion relative. Besso was right. Einstein wrote to Freundlich for help: his "mind was in a deep rut", so he hoped that the young astronomer as "a fellow human being with unspoiled brain matter" could tell him what he was doing wrong. Freundlich could not help him.

The problem, Einstein soon realized, lay with the Entwurf field equations. Worried that Hilbert might beat him to the punch, Einstein rushed new equations into print in early November 1915, modifying them the following week and again two weeks later in subsequent papers submitted to the Prussian Academy. The field equations were generally covariant at last.

In the first November paper, Einstein wrote that the theory was "a real triumph" of the mathematics of Carl Friedrich Gauss and Bernhard Riemann. He recalled in this paper that he and Grossmann had considered the same equations before, and suggested that if only they had allowed themselves to be guided by pure mathematics rather than physics, they would never have accepted equations of limited covariance in the first place.

Other passages in the first November paper, however, as well as his other papers and correspondence in 1913–15, tell a different story. It was thanks to the elaboration of the Entwurf theory, with the help of

**"Worried that Hilbert might beat him to the punch, Einstein rushed new equations into print."**





ETH-BIBLIOTHEK ZÜRICH, BILDARCHIV

ETH Zurich, where Einstein met friends with whom he worked on general relativity.

Grossmann, Besso, Nordström and Fokker, that Einstein saw how to solve the problems with the physical interpretation of these equations that had previously defeated him.

In setting out the generally covariant field equations in the second and fourth papers, he made no mention of the hole argument. Only when Besso and Ehrenfest pressed him a few weeks after the final paper, dated 25 November, did Einstein find a way out of this bind — by realizing that only coincident events and not coordinates have physical meaning. Besso had suggested a similar escape two years earlier, which Einstein had brusquely rejected<sup>2</sup>.

In his third November paper, Einstein returned to the perihelion motion of Mercury. Inserting the astronomical data supplied by Freundlich into the formula he derived using his new theory, Einstein arrived at the result of 43" per century and could thus fully account for the difference between Newtonian theory and observation. "Congratulations on conquering the perihelion motion," Hilbert wrote to him on 19 November. "If I could calculate as fast as you can," he quipped, "the hydrogen atom would have to bring a note from home to be excused for not radiating."

Einstein kept quiet on why he had been able to do the calculations so fast. They were minor variations on the ones he had done with Besso in 1913. He probably enjoyed giving Hilbert a taste of his own medicine: in a letter to Ehrenfest written in May 1916, Einstein characterized Hilbert's style as "creating the impression of being superhuman by obfuscating one's methods".

Einstein emphasized that his general theory of relativity built on the work of Gauss and Riemann, giants of the

mathematical world. But it also built on the work of towering figures in physics, such as Maxwell and Lorentz, and on the work of researchers of lesser stature, notably Grossmann, Besso, Freundlich, Kottler, Nordström and Fokker. As with many other major breakthroughs in the history of science, Einstein was standing on the shoulders of many scientists, not just the proverbial giants<sup>4</sup>. ■

**Michel Janssen** is professor, Program in the History of Science, Technology, and Medicine, University of Minnesota, Minneapolis, USA. **Jürgen Renn** is director at the Max Planck Institute for the History of Science, Berlin, Germany.  
e-mail: [janss011@umn.edu](mailto:janss011@umn.edu);  
[renn@mpiwg-berlin.mpg.de](mailto:renn@mpiwg-berlin.mpg.de)

1. Stachel, J. et al. (eds) *The Collected Papers of Albert Einstein* (Princeton Univ. Press, 1987–2015).
2. Renn, J. (ed.) *The Genesis of General Relativity* Vol. 2 819–830 (Springer, 2007).
3. Gutfreund, H. & Renn, J. *The Road to Relativity* (Princeton Univ. Press, 2015).
4. Renn J. *Auf den Schultern von Riesen und Zwergen: Einsteins unvollendete Revolution* (Wiley VCH, 2006).
5. Janssen, M. & Lehner, C. (eds) *The Cambridge Companion to Einstein* (Cambridge Univ. Press, 2014).
6. Sauer, T. Marcel Grossmann and His Contribution to the General Theory of Relativity. *Proceedings of the 13th Marcel Grossmann Meeting* 456–503 (World Scientific, 2015).
7. Pais, A. 'Subtle is the Lord ...' *The Science and the Life of Albert Einstein* 212 (Oxford Univ. Press, 1982).
8. Clark, R. W. *Einstein: The Life and Times* 156 (Knopf, 1971).
9. Norton, J. D. 'The Hole Argument' *The Stanford Encyclopedia of Philosophy* (ed. Zalta, E. N.) (Fall 2015 Edition); available at <http://plato.stanford.edu>
10. Crellin, J. *Einstein's Jury: The Race to Test Relativity* (Princeton Univ. Press, 2006).

**CORRECTION**

The Comment article 'Einstein was no lone genius' (M. Janssen and J. Renn *Nature* **527**, 298–300; 2015) wrongly stated the dates during which Albert Einstein studied at the Swiss Federal Polytechnical School in Zurich. He was there between 1896 and 1900.





Charles L. Dodgson, better known as Lewis Carroll, in a self-portrait from the 1880s.

## MATHEMATICS

# Logic and Lewis Carroll

As *Alice's Adventures in Wonderland* reaches 150, **Francine Abeles** surveys its creator's wide-ranging legacy.

In 1855, Charles L. Dodgson became the mathematical lecturer at Christ Church College in the University of Oxford, UK. His job was to prepare Christ Church men (for it was all men) to pass examinations in mathematics. Dodgson (1832–98) would go on to publish *Alice's Adventures in Wonderland* (1865) and *Through the Looking-Glass* (1871) under the pen name Lewis Carroll, but he also produced many pamphlets and ten books on mathematical topics.

In some of these, he exhibited unusual methods — for rapid arithmetic, for example. Others featured innovative ideas that foreshadowed developments in the twentieth century, for instance in voting theory. All but two of these books were published by Macmillan (until this year, the parent company of this journal's publisher). Macmillan co-founder Alexander Macmillan was Dodgson's trusted publisher and friend for 35 years (see [go.nature.com/9q8oqe](http://go.nature.com/9q8oqe)).

What unifies Carroll's oeuvre is the wit and colour apparent in the manifestations of his wide-ranging mathematical interests, particularly in geometry and logic. The *Alice* books contain many supreme examples. The "Mad Tea-Party", for instance, has the Hare, Hatter, Dormouse and Alice circling around static place settings like numbers on a circle, as in a modular system, rather than in a line. Carroll developed the earliest modern use of today's 'logic trees', a graphical technique for determining the validity of complex arguments that he called the 'method of trees'. This was a step towards automated approaches to solving multiple connected problems of logic. True to form, the puzzles that Carroll solves with his trees are given quirky names — "The Problem of Grocers on Bicycles", "The Pigs and Balloons Problem".

The ten sections of Carroll's book of droll mathematical stories, *A Tangled Tale*, first appeared between 1880 and 1885 as a serial in the popular magazine *The Monthly Packet*. Carroll dubbed each part a 'knot' to signify the difficulty of the one, two or three problems it featured. In the following issue of the magazine, he would summarize the puzzle, solve it and comment on the solutions he had received from readers, often amusingly presented (see <http://www.onlinemathlearning.com/tangled-tale.html>). *A Tangled Tale* became a favourite of Josiah Willard Gibbs (1839–1903), the applied mathematician and physical chemist praised by Albert Einstein as "the greatest mind in American history".

Carroll believed that beyond their entertainment value, mental recreations such as games and logic puzzles

➔ **NATURE.COM**

To purchase an ebook on Lewis Carroll as logic puzzler, see [go.nature.com/akxwma](http://go.nature.com/akxwma)

conferred a sense of power on the solver. This, he felt, enabled them to analyse any subject clearly and, most important, to detect and unravel fallacies. In this vein, Carroll puns about other knots in *Alice's Adventures in Wonderland*. In Chapter 3, for instance, the Mouse responds to Alice's comments that he had got to the fifth bend in his tale (which appears on the page as a serpentine, tail-shaped paragraph) by crying, "I had not!" Carroll's ever-curious adventurer misunderstands amusingly: "A knot!" said Alice, always ready to make herself useful, and looking anxiously at her. "Oh, do let me help to undo it!"

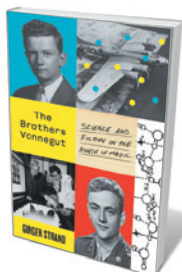
Carroll was, of course, a devotee of word-play, as almost any page of *Alice's Adventures* and *A Tangled Tale* reveals. A fan of acrostics, Carroll dedicated the latter — published in book form in 1885 — to his friend and pupil, the 19-year-old Edith Rix, in the form of a poem that spells her name out in the second letter of each line:

*Beloved Pupil! Tamed by Thee,  
Addish=, Subtrac=, Multiplica=tion,  
Division, Fractions, Rule of Three,  
Attest thy deft manipulation!*

*Then onward! Let the voice of fame  
From Age to Age repeat thy story,  
Till thy hast won thyself a name  
Exceeding even Euclid's glory!*

In the last decades of his life, Carroll published three mathematical pieces in *Nature*. The first, on a method for finding the day of the week for any date (L. Carroll *Nature* 35, 517; 1887), reflects the ▶

## Books in brief



### The Brothers Vonnegut: Science and Fiction in the House of Magic

Ginger Strand FARRAR, STRAUS AND GIROUX (2015)  
Kurt Vonnegut, beloved troublemaker and science-fiction novelist, famously studied chemistry — but it was his brother Bernie who shone in the field. In this engrossing cultural history, Ginger Strand traces the brothers' intellectual development during the Second World War and its chill aftermath. Military interest led Bernie to research silver iodide as a trigger for cloud seeding at General Electric, and Kurt's horrifying experiences in combat inspired his inimitable fiction. Strand shows how both men, by calling in different ways for progress to be decoupled from conflict, revealed a rare integrity.



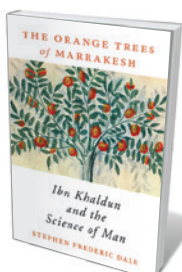
### Thunder & Lightning: Weather Past, Present, Future

Lauren Redniss RANDOM HOUSE (2015)  
Writer and artist Lauren Redniss's *Radioactive* (It Books, 2010; see G. Frazzetto *Nature* 469, 29; 2011) was a beautiful tour de force, meshing superb illustrations with an original telling of the lives of Marie and Pierre Curie. Now, in another aesthetically charged and deeply researched account, Redniss takes on meteorology. Here are phenomena from fog to cyclones; cloud types (a series of nebulous 'portraits'); the sensory appreciation of weather, such as Benjamin Franklin's air bathing, or snowfall's "muffled quietude" — and more. A wild rainstorm of a book, pelting the reader with ideas and inspiration.



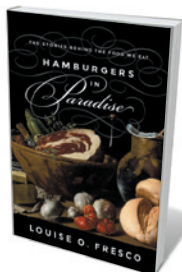
### Fallen Glory: The Lives and Deaths of Twenty Lost Buildings from the Tower of Babel to the Twin Towers

James Crawford OLD STREET (2015)  
This multiple biography of vanished monoliths is itself monolithic, wending its way from Iraq to Manhattan and beyond. Standouts in the narratives built by writer James Crawford include the Tower of Babel, looming in manifestations from Mesopotamian Emperor Nebuchadnezzar's vast ziggurat Etemenanki to Pieter Bruegel the Elder's exquisite and disturbing 1563 painting. The Bastille, Roman Forum, Berlin Wall — all eloquently fall, along with New York's Twin Towers in the hideous events of 2001. My only quibble? No index.



### The Orange Trees of Marrakesh: Ibn Khaldun and the Science of Man

Stephen Frederic Dale HARVARD UNIVERSITY PRESS (2015)  
Six centuries ago, a Tunisian scholar created a new mirror for humankind. In his masterwork *Muqaddimah*, Ibn Khaldun (1332–1406) became the first person to approach history scientifically, by analysing social, economic and political evidence to reveal cycles of societal change. In this sober study, historian Stephen Frederic Dale argues that Ibn Khaldun's work is a key milestone on the road from Greek to Enlightenment thought, chiming with the radical reasoning of philosophers such as Montesquieu and Adam Smith.



### Hamburgers in Paradise: The Stories Behind the Food We Eat

Louise O. Fresco (transl. Liz Waters) PRINCETON UNIVERSITY PRESS (2015)  
Behind the whimsical title is a serious cultural history of food, newly translated from Dutch. Plant scientist Louise Fresco, a former assistant director-general of the Food and Agriculture Organization of the United Nations, argues that the trope of paradise as effortless abundance permeates humanity's tortured relationship with the edible. Her comprehensive trawl through biotechnology, supply chains and more concludes that — given more research and effort — a real paradise of plenty is within reach. [Barbara Kiser](#)

ILLUSTRATION BY JOHN TENNIEL FROM ALICE'S ADVENTURES IN WONDERLAND (MACMILLAN, 1995)



Alice's circular conversation with the Caterpillar is a gem of semantic wordplay.



► calendar problems of the time: to obtain information on future days and dates, you had to consult an almanac. Carroll found mental calculation methods gripping. Introducing the piece, he wrote, “I am not a rapid computer myself”; yet noted that he could do ten such problems in less than four minutes. His rule uses four integer calculations: two for the year, the third for the month and the last for the day.

In an era before calculators, standard arithmetic processes were onerous and prone to error. Carroll (writing this time under his real name) summarized his work on simplifying ordinary arithmetical calculations in his second piece in *Nature*, ‘Brief Method of Dividing a Given Number by 9 or 11’

**“Carroll created a vivid tapestry of work, presaging in many ways developments in the twentieth century and beyond.”**

(C. L. Dodgson *Nature* 56, 565–566; 1897) which also included division by 13, 17, 19 and 41, as well as by numbers within 10 from a power of 10, either way. The third piece, ‘Abridged Long Division’ (C. L. Dodgson *Nature* 57, 269–271; 1898), by his own admission, uses ideas put forth by others that he improved on, particularly an accuracy test. However, this paper has implications for modern computing in its emphasis on minimizing the number of steps in an algorithm.

Carroll did not influence his contemporary colleagues in the development of mathematical ideas. However, posthumously, beginning in the last half of the twentieth century, his contributions to voting theory were uncovered in three papers written between 1874 and 1876. The third, ‘A Method of Taking Votes on More Than Two Issues’, is the most important. Carroll was the first to create a voting method that would achieve biproportional representation — that is, proportionality with respect both to the population in the districts and to



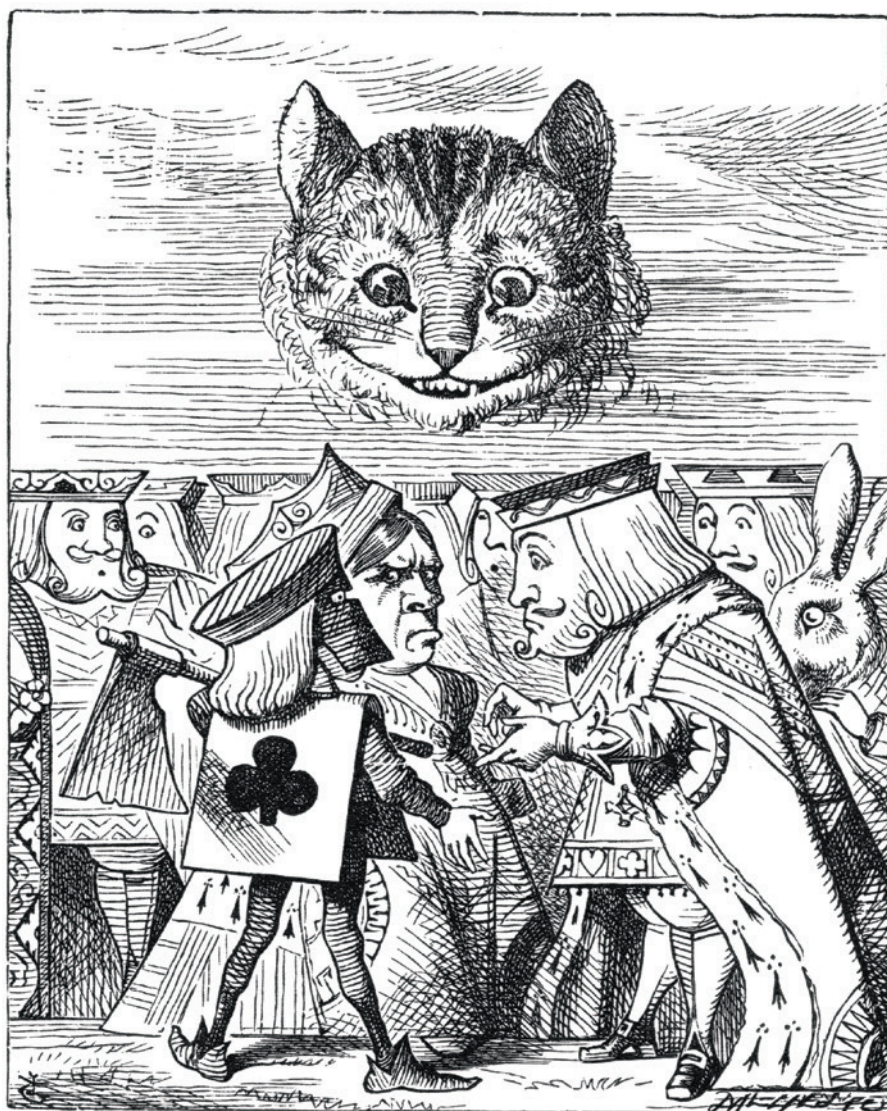
Carroll's publisher, Alexander Macmillan.

the apportionment of seats to the political parties in the legislature. Despite Carroll's friendship with Lord Salisbury, the UK prime minister at the time, it was not applied for political reasons. (Today, the European Parliament uses a form of proportional representation.)

Carroll's work in logic, notably the unpublished second part of his book *Symbolic Logic*, foreshadowed results that appeared about 100 years later. This long-lost section, which contains the method of trees, was described by philosopher W. W. Bartley (*Sci. Am.* 227, 38–46; 1972). Carroll's book on linear algebra (*An Elementary Treatise on Determinants with Their Application to Simultaneous Linear Equations and Algebraic Geometry*, 1867) is also groundbreaking. His ‘condensation’ method for computing determinants sparked research that led to a formulation of the alternating sign matrix conjecture by David Robbins and Howard Rumsey in the 1980s. And his 1895 ‘What the Tortoise Said to Achilles’, a logic problem he published in the philosophical journal *Mind*, remains unsolved. In 1858, he was the first to create a cypher in matrix form based on a non-standard (modular) arithmetic; it was published more than 100 years later.

As a mathematician, logician, writer and innovative photographer, Carroll created a vivid tapestry of work, knotted, twisted and multistranded, and presaging in many ways developments in the twentieth century and beyond. Yet for all the complexity and playfulness of this master gamester's body of work — from voting theory to his great creation, Alice, on her long, strange journeys towards identity and maturity — his underlying concerns were fairness, certainty and truth. ■

**Francine F. Abeles** is professor emerita of mathematics and computer science at Kean University in Union, New Jersey.  
e-mail: [fabeles@kean.edu](mailto:fabeles@kean.edu)



A love of puzzles is clear in the call to behead the bodyless Cheshire Cat: what, exactly, would you behead?

CHRONICLE/ALAMY

ILLUSTRATION BY JOHN TENNIEL FROM ALICE'S ADVENTURES IN WONDERLAND (MACMILLAN, 1995)



# Correspondence

## Peat fires: emissions likely to worsen

The horrific haze from Indonesia's forest and peatland fires, started deliberately to clear land for planting and made worse by drought, has become a global crisis. Indonesia's government could stop this annual catastrophe, but it so far seems to lack the political will to do so.

In the past decade, Indonesia has destroyed its forests faster than any other nation (see [go.nature.com/b9rhxz](http://go.nature.com/b9rhxz)). By one estimate, daily carbon emissions from its forest and peatland fires now exceed those from the entire US economy (see [go.nature.com/hpworu](http://go.nature.com/hpworu)).

The situation is likely to worsen: Indonesia and Malaysia are planning to set up a Council of Palm Oil Producing Countries. This intends to force major forest-exploiting corporations to relax their zero-deforestation pledges (see [go.nature.com/agvbhn](http://go.nature.com/agvbhn)). Oil-palm expansion is one of the biggest drivers of peatland and forest destruction.

Localized actions and belated half-measures by the Indonesian government are no longer enough. Aided by the global community, it must ban fires in peatlands and native forests; declare a moratorium on clearing peatlands; restore water to degraded peatlands; and create financial incentives for provinces to reduce deforestation.

**Susan G. Laurance, William F. Laurance** *James Cook University, Cairns, Queensland, Australia.*  
[susan.laurance@jcu.edu.au](mailto:susan.laurance@jcu.edu.au)

## Peat fires: consumers to help beat them out

Southeast Asia's choking air pollution continues unabated, fuelled by the burning of peat swamps for agriculture. The issue flies in the face of long-standing regional agreements on land clearance by governments in the Association of Southeast Asian Nations, and is at last galvanizing

non-governmental organizations (NGOs), banks and businesses into action against the companies responsible.

Singapore's punitive Transboundary Haze Pollution Act has met with some success (see, for example, J. H. S. Lee *et al. Environ. Sci. Policy* **55**, 87–95; 2016). The Singapore Environment Council, an NGO, has suspended environmental certification of paper-pulp companies that might be connected with the fires. This has prompted some supermarket chains in Singapore to stop selling products containing raw materials from these companies, and banks are reviewing their policies for lending to them. Suspension could prompt companies to become more sustainable and to consider setting aside undeveloped peat-swamp forests for conservation.

Consumers should back this drive for corporate environmental accountability by using publicly available resources (see, for example, [www.ethical.org.au](http://www.ethical.org.au)) to ensure that their product choices do not result in peat clearance.

**Lahiru S. Wijedasa** *National University of Singapore; and Conservation Links, Singapore.*  
**Mary Rose C. Posa** *National University of Singapore, Singapore.*  
**Gopalasamy R. Clements** *University of Malaysia Terengganu, Kuala Terengganu, Malaysia.*  
[lahirux@gmail.com](mailto:lahirux@gmail.com)

## Time for Russia to tap renewables

Russia's big territory and coastline are potentially huge sources of renewable energy from sun, wind, waves, tides and currents, but about 91% of the country's energy still comes from fossil fuels. This must be urgently rectified if Russia is to honour its pledge, made ahead of this month's climate summit in Paris, to reduce its carbon emissions by 25–30% relative to 1990 levels by 2030.

Russia's carbon emissions have been increasing since 1998. Only

3.2% of its total primary energy supply came from renewables in 2013 (nuclear accounts for the rest; see [www.iea.org/statistics](http://www.iea.org/statistics)). This compares poorly with industrial nations such as Brazil (40%), Sweden (35.7%), India (26.4%), Canada (18.6%) China (11%) and the United States (6.8%).

Russian environmental legislation is taking small, ongoing steps to protect its natural resources, clean up polluted areas, control air and water quality and advance green industrial technologies. Environmental penalties for pollution and illegal use of natural resources have increased sharply. More investment in renewable energy will help to protect Russia's natural environment.

**Alexander Gorobets** *Sevastopol, Russia.*  
[alex-gorobets@mail.ru](mailto:alex-gorobets@mail.ru)

## Cannabis: monitor policy changes

Policy changes that increasingly permit the medical and recreational use of cannabis have important implications for society and drug policies overall (see *Nature* **524**, 280–283; 2015 and *Nature* **525**, S1–S18; 2015). There is an urgent need to set up collaborative international monitoring of the effects of these changes in different countries to achieve a meaningful evaluation of their impact.

Monitoring should include differences in arrest numbers, imprisonment, public-health and social impact, supply and economic analyses, as well as co-dependence and substitution effects on the use of other drugs, alcohol and tobacco products ([go.nature.com/jr5lgt](http://go.nature.com/jr5lgt)).

The effects of cannabis policy changes might be influenced by, for example, a legal age limit, accurate labelling of contents and potency, restrictions on advertising, and law-enforcement practices.

**Lucas Wiessing\*** *European Monitoring Centre for Drugs and*

*Drug Addiction, Lisbon, Portugal.*  
[lucas.wiessing@emcdda.europa.eu](mailto:lucas.wiessing@emcdda.europa.eu)  
\*On behalf of 5 correspondents (see [go.nature.com/tn7ylb](http://go.nature.com/tn7ylb) for full list).

## Cannabis: debated schizophrenia link

In our view, Matthew Hill's arguments against a causal link between cannabis use and schizophrenia fail to clinch this debate (*Nature* **525**, S14; 2015).

His contention that the increased societal use of cannabis over time is not reflected in increased rates of schizophrenia has been tested only once to our knowledge — and that study came to the opposite conclusion (J. Boydell *et al. Psychol. Med.* **36**, 1441–1446; 2006). In multifactorial conditions such as schizophrenia, an increase in one risk factor is not necessarily balanced by a decrease in another. Deaths from cardiac disease are declining in many countries despite increased obesity, but that does not mean that obesity is unrelated to cardiac disease.

Hill misinterprets our review of cannabis use by people with psychosis (A. Kolliakou *et al. Intl J. Dev. Neurosci.* **29**, 335–346; 2011). Contrary to his inference that this group self-medicates to mitigate negative symptoms, we found that the most commonly reported use by these individuals was purely recreational.

He suggests that cannabis is a risk factor only for those with a genetic predisposition to schizophrenia (but see M. Di Forti *et al. Biol. Psychiatry* **72**, 811–816; 2012). Another explanation could be that some of the genes associated with a proclivity for cannabis smoking also show up among those who are predisposed to schizophrenia, because a genetic tendency for the habit could in turn increase the risk of schizophrenia.

**Matthew Large** *University of New South Wales, Australia.*  
**Marta Di Forti, Robin Murray** *Kings College London, UK.*  
[mmbl@bigpond.com](mailto:mmbl@bigpond.com)



# Richard Heck

## (1931–2015)

Organic chemist who won a Nobel for palladium catalysis.

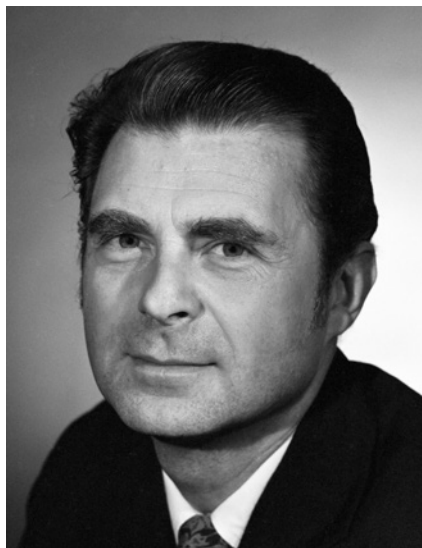
“Do something with transition metals,” quipped the research director of the Hercules Powder Company to Richard Heck in 1958, two years after he had joined. After consulting with Pat Henry, an organometallic chemist across the hall, and with some bold thinking, Heck discovered a new method for clicking together carbon atoms in a single step (R. F. Heck *Synlett* **18**, 2855–2860; 2006).

The carbon–carbon bond is a prerequisite for life: nature catalyses it with enzymes; Heck used palladium. It was the dawn of a new era in organic synthesis, the field dedicated to constructing a vast array of compounds, from simple building blocks to bewildering mega-atom frameworks. Heck had taken the first step on the path that led him to the 2010 Nobel Prize in Chemistry.

Ask an organic chemist today about products of the Heck reaction and they’ll name smartphone displays, sunscreens, perfumes, pesticides and medicines. One example is the over-the-counter pill naproxen for pain, fever, stiffness and inflammation. A biologist will recognize the reaction as the basis for the coupling of fluorescent dyes to DNA bases, allowing the automation of DNA sequencing and the elucidation of the human genome.

Richard Heck died on 9 October 2015 in Manila. He was born in Springfield, Massachusetts, on 15 August 1931. Aged eight, he moved with his parents, both professional dancers, to Los Angeles in California. His interest in chemistry was stirred by the vivid colours and abundant fragrances of flowers in the vacant lot near their home. Following a PhD with the prominent physical organic chemist Saul Winstein at the University of California, Los Angeles, and a postdoctoral fellowship with the future Nobel laureate Vladimir Prelog at the Swiss Federal Institute of Technology in Zurich, the 25-year-old Heck joined Hercules (now Ashland) in Wilmington, Delaware, in 1956.

After two years of working on the development of a commercial process for producing polyethylene using the newly discovered Ziegler–Natta catalysts, Heck was given the fateful mission by research director David Breslow. “They left us alone to try anything we want,” Heck later said. So, appreciating that discovery proceeds stepwise from scattered observations in the literature, Heck studied the alkene hydroformylation reaction. He proposed the first correct



mechanism for a reaction catalysed by a transition metal. It illuminated many other unexplained organometallic reactions, and produced a rich harvest of new cobalt organometallic chemistry.

The hydroformylation technology is currently used to produce 6.8 million tonnes of basic carbon building blocks (alcohols and aldehydes) each year for the synthesis of everyday materials. Perhaps less known today are Heck’s other forays into cobalt carbonyl chemistry to establish reactions with a variety of organic molecules (carbon monoxide, alkenes, dienes, epoxides and ketones). With, as he put it, “no immediate ideas” on how to employ this chemistry profitably for Hercules, he took a new direction.

In 1968, the chemical community was astounded by Heck’s flurry of seven consecutive single-author papers in the *Journal of the American Chemical Society*. In hindsight, these heralded the innovative work to follow when he moved down the road to the University of Delaware in 1971. The next year, Heck’s seminal paper appeared (R. F. Heck & J. P. Nolley *J. Org. Chem.* **37**, 2320–2322; 1972). With characteristic generosity, he begins by acknowledging Tsutomu Mizoroki for preceding his discovery. He goes on: “We have independently discovered this reaction and find that it can be carried out under much more convenient laboratory conditions.”

The discovery, now also widely known as the Mizoroki–Heck reaction, involves the palladium catalyst slipping into a carbon–halogen bond to give a fleeting species.

This intermediate clutches another reactive molecule to form, after more contortions, a product in which a new carbon–carbon bond is established. Although this paper led to the Nobel prize in 2010 (with Ei-ichi Negishi and Akira Suzuki), it lay mostly unappreciated in the literature. Heck continued on his innovative path with two further publications, in 1975, which revealed two other new ways of carbon–carbon bond formation that are harbingers of the Sonogashira and Suzuki–Miyaura cross-coupling reactions. “I reported the copperless Sonogashira,” Dick Heck said to me unpretentiously.

During this prodigious run of groundbreaking research, he and his co-workers established two other mainstays of the synthetic chemists’ toolbox: palladium-catalysed carbonylation of aryl halides and transfer hydrogenation with formate as a reductant.

Today, undergraduates learn the Heck reaction in class and laboratory; industrial chemists practise it to make tonnes of drugs against asthma, diabetes and AIDS, among others. Thus, Heck’s work may be considered as a forerunner of a cornucopia of transition-metal-catalysed technologies that are in operation worldwide.

In 1989, he retired to Florida with his Filipino wife Socorro Nardo. In 2006, I invited him to Queen’s University, Canada, to follow up on his cobalt work. Students overcame their awe to work alongside him, striving to arrive before his customary 8 a.m. start. Seventeen years after retirement and 45 years after opening the door to organocobalt chemistry, Dick entered the lab again, prepared the complexes and measured their infrared spectra (“you get all of the information you need”). With the assistance of a postdoc, he obtained nuclear magnetic resonance and high-resolution mass spectrometry data, interpreted them, and took the next step.

Dick returned later in 2006 to the Philippines with Socorro. A handwritten letter to me noted that he had returned to his two passions: “I have some room to grow orchids again so I will have something to do,” and “Have you found anyone to work with cobalt carbonyl?” He had drawn chemical structures of potential next steps along the palladium-paved path that he had established. ■

**Victor Snieckus** is at Queen’s University, Ontario, Canada. Richard Heck worked in his laboratories in 2006.  
e-mail: victor.snieckus@chem.queensu.ca

UNIV. DELAWARE

## IMAGING TECHNIQUES

# Extra dimension for bone analysis

A combination of two techniques — computed tomography and small-angle X-ray scattering — and serious computing power have enabled multi-scale, three-dimensional analysis of bone and tooth tissue. [SEE LETTERS P.349 & P.353](#)

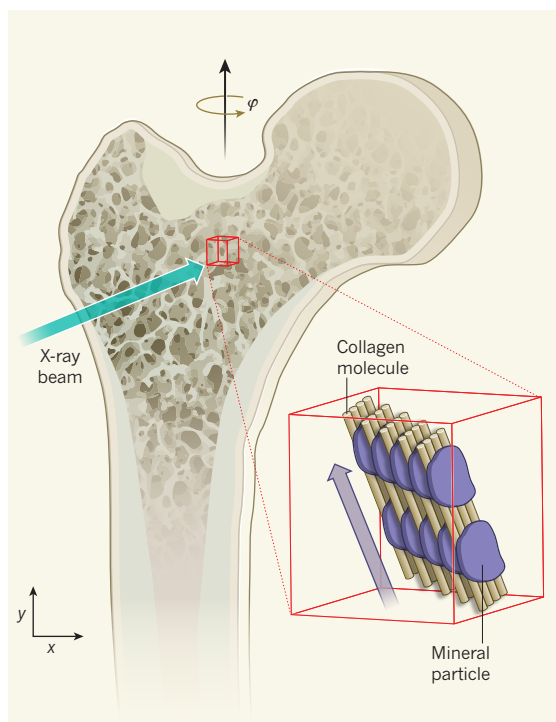
PETER FRATZL

In this issue, two papers (by Liebi *et al.*<sup>1</sup> and by Schaff *et al.*<sup>2</sup>) describe different approaches to visualizing three-dimensional bone and tooth structures on both macroscopic and nanometre scales. The methods derive from computed tomography, a well-established imaging technique that can yield 3D pictures of bones, but which quantifies just one scalar parameter, such as mineral density, in each voxel (3D pixel) of the image. Using the new techniques, each voxel contains information on both the orientation and the size of mineral particles in the bone or tooth specimen.

The strength of bone is determined by its structure on all scales, particularly by local variation in the size, amount and orientation of bone mineral particles<sup>3</sup>. Formed from calcium phosphate, these mineral platelets are just a few nanometres thick and are embedded in a matrix of collagen molecules (Fig. 1, inset). Because bone tissue is constantly remodelled and adapted, the orientation of mineralized collagen fibrils varies in a complex fashion throughout the tissue<sup>4</sup>.

Although transmission electron microscopy provides sufficient resolution to visualize the mineral particles, it does not allow them to be mapped over millimetre distances in complete bone sections. For this reason, a technique called small-angle X-ray scattering (SAXS) has been used in a 'scanning' mode since the 1990s to study bone taken from patient biopsies and animals<sup>5,6</sup>. In this method, a bone specimen is moved in two dimensions (defined by the  $x$  and  $y$  axes in Fig. 1) across a narrow X-ray beam, and at each position of this scan a SAXS pattern is collected. This allows the simultaneous visualization of two different scales: the structure, particularly the orientation, of the mineralized collagen fibrils in every pixel of the scan; and the variation of fibril orientation across a macroscopic specimen.

When a 2D X-ray detector is used, as in the early implementations of scanning SAXS, it effectively provides 4D information:



**Figure 1 | Principle of small-angle X-ray scattering (SAXS) tomography.** Two papers<sup>1,2</sup> report methods for SAXS tomography that enable multi-scale analysis of bone and tooth structures. A sample (here, a bone specimen) is scanned by a narrow X-ray beam in the  $x$  and  $y$  directions for various rotation angles  $\varphi$  around an axis through the sample. The scans provide 2D maps that are combined to produce a 3D reconstruction of the bone structure. Unlike conventional methods for visualizing bones, SAXS tomography provides both a 3D macroscopic map of the sample and 3D information on the nanoscale mineral particles embedded in the collagen molecules of bone (such as their predominant direction of alignment; purple arrow in inset).

a macroscopic 2D map of the bone tissue and 2D information about the structure of the mineralized collagen fibrils in every pixel of the map<sup>7</sup>. It is straightforward to extend this to five dimensions by rotating the bone section around an axis to collect 3D information on the mineralized collagen fibrils in each voxel, while the mapping stays 2D<sup>8</sup>.

It would be desirable to extend this to six dimensions by enabling 3D macroscopic mapping of the specimen. But the reconstruction of such 6D data is a formidable numerical challenge and can be solved only through

the use of certain approximations and substantial computing power. In conventional computed tomography, volumes are reconstructed from 2D images taken in the  $x$ - $y$  plane for many rotation angles  $\varphi$  around a given axis (Fig. 1). For SAXS tomography, 2D SAXS data must be collected at each position of the  $x$ - $y$  plane and at many rotation angles, not just around one axis but around many axes. This results in a huge number of measurements and requires massive computational effort.

To make such efforts tractable, Liebi *et al.* (page 349) take advantage of certain symmetries in the mineralized collagen fibrils of bone. More specifically, they assume that the arrangement of mineral particles is rotationally symmetrical — it looks the same after a certain amount of rotation — around the direction in which collagen molecules align in each fibril. This symmetry imposes constraints on the SAXS patterns that aid the reconstruction of 3D images using a procedure sometimes called tensor tomography.

Schaff and colleagues' approach (page 353) assumes that the SAXS signal varies slowly with rotation angle. This means that fewer rotation angles are needed during data collection, because data values between rotation angles can be interpolated. The authors used their technique to study dentine in the interior of a tooth. Both techniques<sup>1,2</sup> produce 3D pictures of macroscopic specimens, in which each voxel contains an arrow that represents the direction of the predominant orientation of the mineralized collagen fibrils in the voxel.

A previous attempt<sup>9</sup> to develop 3D scanning-SAXS tomography was devised for materials with structures that have rotational symmetry around one axis, such as composite materials based on parallel fibres. This approach is similar to that of Liebi and colleagues, but it assumes that all the fibres point in the same direction throughout the specimen, which is not true for bone. Another recently reported approach<sup>10</sup> applies normal scanning SAXS to a series of consecutive thin bone sections cut



from one block; by assembling the resulting series of scanning SAXS data, a full 3D image was reconstructed. The obvious disadvantage of that technique compared with the currently reported ones is that the bone sample was destroyed.

Both Liebi *et al.* and Schaff *et al.* collected more than one million SAXS patterns for their reconstructions — equating to terabytes of data, an impressive amount. Recording them using a synchrotron X-ray source took about a day, and the computations needed for one tomogram required several days of computing time. This is manageable for proof of principle of the techniques, but would be prohibitive for

clinical studies, for which many tomograms would be required.

Nevertheless, the feasibility of the techniques has been demonstrated. With the progress currently foreseeable in the development of X-ray sources, detectors and computing power, one can therefore expect SAXS tomography to become an important tool for the analysis of bone, dentine and other mineralized tissues in biological and medical studies. ■

**Peter Fratzl** is in the Department of Biomaterials, Max Planck Institute of Colloids and Interfaces, Potsdam 14424, Germany. e-mail: fratzl@mpikg.mpg.de

1. Liebi, M. *et al.* *Nature* **527**, 349–352 (2015).
2. Schaff, F. *et al.* *Nature* **527**, 353–356 (2015).
3. Weiner, S. & Wagner, H. D. *Annu. Rev. Mater. Sci.* **28**, 271–298 (1998).
4. Fratzl, P. & Weinkamer, R. *Prog. Mater. Sci.* **52**, 1263–1334 (2007).
5. Fratzl, P., Jakob, H. F., Rinnerthaler, S., Roschger, P. & Klaushofer, K. *J. Appl. Cryst.* **30**, 765–769 (1997).
6. Rinnerthaler, S. *et al.* *Calcif. Tissue Int.* **64**, 422–429 (1999).
7. Pabisch, S., Wagermaier, W., Zander, T., Li, C. & Fratzl, P. *Meth. Enzymol.* **532**, 391–413 (2013).
8. Jaschouz, D., Paris, O., Roschger, P., Hwang, H.-S. & Fratzl, P. *J. Appl. Cryst.* **36**, 494–498 (2003).
9. Stribeck, N. *et al.* *Macromolecules* **41**, 7637–7647 (2008).
10. Georgiadis, M. *et al.* *Bone* **71**, 42–52 (2015).

## ANTIBIOTICS

# Homed to the hideout

**Some *Staphylococcus aureus* bacteria are thought to survive standard antibiotic treatment by ‘hiding’ in host cells. But an antibody–antibiotic conjugate has been developed that targets these bacteria in mouse models. [SEE ARTICLE P.323](#)**

WOLF-DIETRICH HARDT

The pathogenic bacterium *Staphylococcus aureus* causes thousands of deaths each year. Therapy is sometimes unsuccessful, partly because antibiotic-resistance genes are spreading worldwide. However, even strains of *S. aureus* that lack resistance genes are often difficult to kill with available antibiotics; it has been suggested that the bacteria ‘hide’ inside host cells. This hypothesis inspired Lehar *et al.*<sup>1</sup>, who, on page 323 of this issue, present a construct in which an antibiotic is linked to an antibody that binds to the pathogen’s surface. Alone, this ‘prodrug’ is inactive, but when prodrug-coated bacteria enter host cells, enzymatic activity releases the antibiotic. In mouse models of *S. aureus* infection, this strategy was strikingly more potent than standard antibiotic treatment.

Antibiotics are a pillar of modern medicine, but they are not effective in all cases. There are at least three explanations for this. First, important pathogenic bacteria, including many *S. aureus* strains, have acquired resistance against standard antibiotics<sup>2</sup>. Second, the pathogen may hide in host sites that cannot be reached by antibiotic molecules, or where the environmental conditions, such as high acidity, render the antibiotics inactive. And third, in certain circumstances, bacteria can switch to a ‘persistent’ lifestyle that makes them insensitive to antibiotics<sup>3</sup>. The switch to persistence is still not completely understood but can occur in some pathogens when they enter host cells.

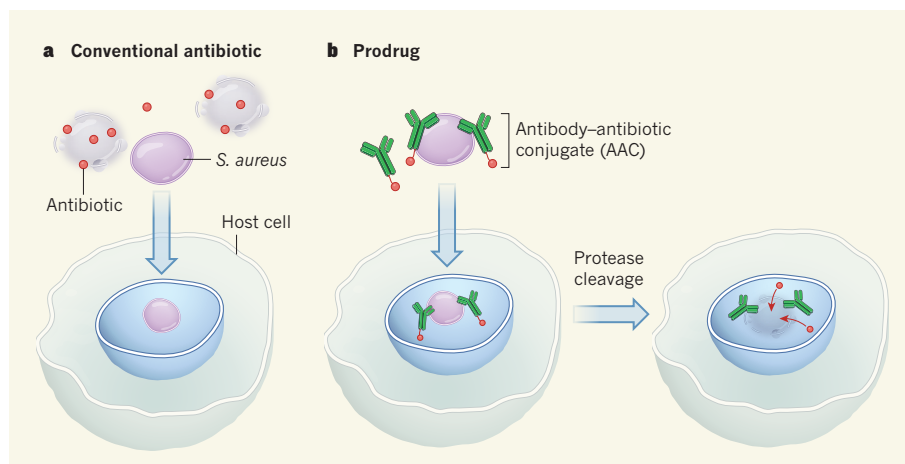
In *S. aureus*, a range of virulence factors manipulate host-cell processes, prohibit efficient immune responses and fuel infections

in wounds, the bloodstream and other sites. In the absence of antibiotic-resistance genes, several different classes of antibiotic are used to treat *S. aureus* infections; rifampicin antibiotics are sometimes employed to target intracellular reservoirs of the bacterium. However, these classic antibiotics are often unable to cure the infection.

Lehar and colleagues speculated that this failure is due to either insufficient accessibility of the antibiotic or insufficient activity against intracellular *S. aureus*. In an attempt to overcome these problems, the researchers

first generated a rifampicin derivative (a ‘rifalogue’) with altered physicochemical properties that gives superior activity against *S. aureus* cells that have switched to a persistent lifestyle. Next, they identified an antibody that tightly binds to sugar structures found on the surface of all *S. aureus* strains tested. Then they covalently joined these two components by using a chemical bridge that can be broken by protease enzymes that are present at the intracellular sites where *S. aureus* is thought to hide out (Fig. 1). Strikingly, in a mouse infection model, this antibody–antibiotic conjugate (AAC) was much more effective at reducing pathogen loads than two conventional antibiotics currently used to treat recalcitrant *S. aureus* infections.

This approach is reminiscent of antibody-targeted prodrug strategies that are currently used in cancer therapy<sup>4</sup>, and these proof-of-principle data suggest that targeted antibiotic delivery is a promising strategy for fighting obstinate intracellular pathogens. It remains to be seen whether AACs are as efficient at



**Figure 1 | Targeted intracellular antibiotic release.** **a**, *Staphylococcus aureus* infections are notoriously difficult to treat. It is thought that this is because the bacteria enter host cells and ‘hide’ in intracellular compartments that conventional antibiotics cannot reach or where they are inactive. **b**, Lehar *et al.*<sup>1</sup> covalently linked an antibiotic derivative, called a rifalogue, to an antibody that binds to components of the *S. aureus* cell wall. This prodrug coats the bacterial cell surface but remains inactive until the bacteria enter the host cell. There, protease enzymes cleave the linker region, releasing the active antibiotic, which then kills the bacteria.

treating bacterial infections in humans as they are in mice, especially in chronically infected patients, who often already have antibodies against *S. aureus*. Such antibodies may shield the bacterial surface from AAC binding, and therefore interfere with the targeting of the prodrug. Moreover, because the antibiotic makes up only around 1% (by mass) of the current construct, the AAC would have to be applied at the equivalent of more than a gram per dose for an adult human patient. This might be improved in the future by replacing the antibody with smaller surface-targeting entities.

Why is the AAC approach so much more effective than standard antibiotics? One reason is that the rifalogue is more efficient than rifampicin at killing persistent *S. aureus* cells. Another is that the kinetics of drug distribution, excretion and inactivation seem to be favourably affected by its fusion to the antibody. Coating bacterial cells with the antibody-bound prodrug may also steer the bacteria to be taken up into intracellular compartments (lysosomes) that have high levels of the enzymes needed to release the antibiotic<sup>5</sup>. And accumulation of

the AAC on the pathogen's surface may cause particularly high local concentrations of the bacteria in the intracellular hideout. It remains to be seen which of these mechanisms account for the *in vivo* potency of the AAC.

Compared with conventional antibiotic therapy, the prodrug approach is likely to reduce both the emergence of antibiotic resistance (by reducing the exposure of other bacteria to the active drug) and the disruption of the body's normal communities of microorganisms. There is still plenty of scope for optimizing the targeting moieties and the chemical bridges<sup>6</sup>. Moreover, the strategy may allow researchers to revisit older antimicrobials that were not developed for therapy because they had unfavourable pharmacokinetics or toxicity. The AAC approach could also expand our arsenal against other notorious intracellular pathogens, such as *Mycobacterium tuberculosis*.

Alternative strategies to tackle the growing problem of antibiotic resistance are also emerging. These include antibiotics that specifically target persistent cells<sup>7</sup>, agents that stimulate the host's antimicrobial defences

to augment antibiotic therapy<sup>8,9</sup>, or harmless 'biocontrol' agents that colonize the host and inhibit pathogen growth<sup>10</sup>. We can hope that such approaches, alongside the AAC strategy presented by Lehar *et al.*, will boost our capacity to treat bacterial infections. ■

**Wolf-Dietrich Hardt** is at the Institute of Microbiology, ETH Zürich, 8093 Zürich, Switzerland.

e-mail: [hardt@micro.biol.ethz.ch](mailto:hardt@micro.biol.ethz.ch)

1. Lehar, S. M. *et al.* *Nature* **527**, 323–328 (2015).
2. World Health Organization. *Antimicrobial Resistance: Global Report on Surveillance* (WHO, 2014).
3. Lewis, K. *Annu. Rev. Microbiol.* **64**, 357–372 (2010).
4. Giang, I., Boland, E. L. & Poon, G. M. K. *AAPS J.* **16**, 899–913 (2014).
5. Joller, N. *et al.* *Proc. Natl Acad. Sci. USA* **107**, 20441–20446 (2010).
6. Yacoby, I. & Benhar, I. *Infect. Disord. Drug Targets* **7**, 221–229 (2007).
7. Conlon, B. P. *et al.* *Nature* **503**, 365–370 (2013).
8. Kaiser, P. *et al.* *PLoS Biol.* **12**, e1001793 (2014).
9. Porte, R. *et al.* *Antimicrob. Agents Chemother.* **59**, 6064–6072 (2015).
10. Iwase, T. *et al.* *Nature* **465**, 346–349 (2010).

This article was published online on 4 November 2015.

## ASTROPHYSICS

# Growing planet brought to light

**Thousands of extrasolar planets have been discovered, but none is a planet in its infancy. Observations have finally been made of a young planet growing in its birthplace — opening the way to many more such discoveries. SEE LETTER P.342**

ZHAOHUAN ZHU

Finding young planets in their birthplaces is extremely challenging, because actively forming planetary systems are far away and obscured by dust. On page 342 of this issue, Sallum *et al.*<sup>1</sup> report the use of a new technique to detect an emission signal from a growing planet. This discovery has far-reaching implications for our understanding of the planet-formation process and of the properties of young planets.

When a star is born, a flat rotating disk of gas and dust forms around it, known as the circumstellar disk. This disk continuously transports dust and gas inward to feed the young star for millions of years, a process known as disk accretion. Planets are thought to form from the leftover material from this disk. Earth, for example, was born in the circumstellar disk surrounding the young Sun 4.6 billion years ago. But little is known about how microscopic dust particles can grow 14 orders of magnitude bigger to become a giant planet within the relatively short lifetime of the

disk. Finding young planets in circumstellar disks should provide important clues about when, where and how young planets are born.

But finding planets is difficult because they are small and dim. The presence of most known planets has therefore been inferred from observations of the stars around which they revolve. For instance, the Kepler satellite<sup>2</sup> has discovered more than 1,000 planets by measuring the tiny dimming of stellar light that occurs when a planet passes in front of its star.

Such methods cannot be used to find young planets around young stars, because such stars are highly active and the light they emit is variable. Most attempts to find young planets use 10-metre-diameter optical telescopes to directly image planets in circumstellar disks. Disks with large cavities are particularly targeted<sup>3</sup>, because such cavities are thought to be opened up by giant planets in orbit around the central star.

In 2012, a protoplanet candidate 1,000 times fainter than its host star was discovered<sup>4</sup> in a system called LkCa 15 (Fig. 1). The central

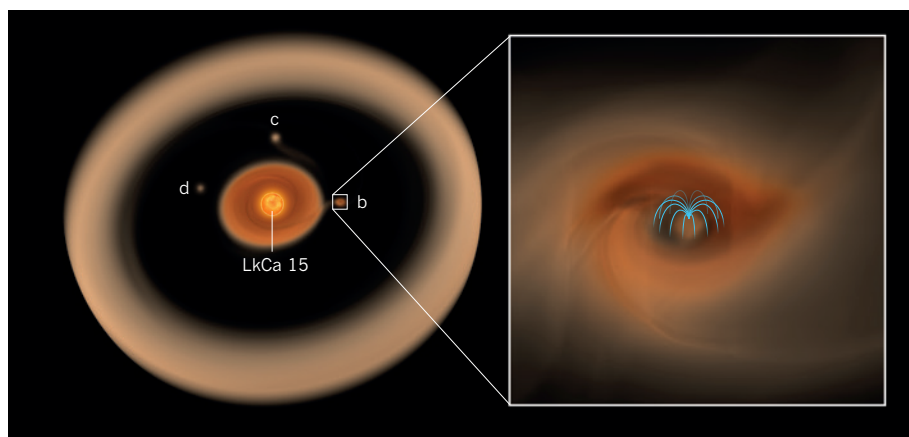
star of this system is similar to our Sun, but is only 2 million years old. The Sun-like star has a circumstellar disk with a cavity 50 astronomical units in radius (1 AU is the distance from Earth to the Sun).

The protoplanet candidate, called LkCa 15 b, resides inside the cavity, 16 AU away from the central star. But the nature of LkCa 15 b is unclear — it is redder than would be expected for a young planet. Several young planet candidates<sup>5,6</sup> have since been discovered in other circumstellar disks, and they are all quite red. This has led to the hypothesis that the detected red objects are young planets with circumplanetary disks. When a planet is born, a rotating circumplanetary disk of gas and dust forms around it, similar to the circumstellar disks around young stars. As the accreting disk feeds the nascent planet, it releases energy and becomes bright. The emission from such a disk should be redder than the planet itself<sup>7</sup>.

In their study, Sallum *et al.* searched for a signature of young planets: H $\alpha$  photons, which are emitted from hydrogen atoms only when a circumplanetary disk accretes onto a planet. If a young planet has strong magnetic fields, the fields form a large magnetosphere around the planet, which can truncate the circumplanetary disk<sup>7,8</sup>. The material in the disk therefore has to follow the planet's magnetic fields to accrete onto the planet. During this accretion process, the magnetosphere can be as hot as 10,000 kelvin, which is what causes hydrogen atoms to emit H $\alpha$  photons<sup>9</sup>.

Although the emission of H $\alpha$  photons has been widely observed when circumstellar disks accrete onto young stars, Sallum and colleagues are the first to directly image accreting





**Figure 1 | Circumplanetary-disk discovery.** The young star LkCa 15 is surrounded by a disk of dust and gas. Sallum *et al.*<sup>1</sup> report that a young planet (LkCa 15 b) is growing in a gap in that circumstellar disk, and that two other potential young planets (LkCa 15 c and d) also reside within the gap. Disks of dust and gas also form around young planets (inset), providing material for them to grow continuously. When material from a circumplanetary disk follows the magnetic fields of young planets (blue curves) to be accreted onto those planets, it produces light known as H $\alpha$  photons. The authors report that LkCa 15 b is an H $\alpha$  emitter. This graphic is based on supercomputer simulations of the gas distribution in the LkCa 15 system; the central star and planets are not shown to scale. (Graphic modified from images provided by Z. Zhu.)

circumplanetary disks around young planets using H $\alpha$  photons. To do this, they used a filter that allows only H $\alpha$  photons to reach their telescope<sup>10</sup>. The authors report that LkCa 15 b is an H $\alpha$  emitter, providing strong evidence that it is a young planet with a circumplanetary disk still accreting onto the planet. Furthermore, they found two other objects inside the cavity of the LkCa 15 system, although these do not seem to be H $\alpha$  emitters. By combining information from observations made over several time periods with data from the initial discovery in 2012, Sallum *et al.* determined the orbits of two of the young planet candidates.

The researchers' discovery provides stringent constraints on planet-formation theories. For example, such theories now have to explain how a giant planet can form 15–16 AU from its star within 2 million years, and still be growing after this time. Another implication of the findings is that a young planet's magnetic fields need to be at least 20 times stronger than Jupiter's current magnetic fields to truncate the accreting circumplanetary disk. This in turn implies that the internal motion of young giant planets is much greater than that of the giant planets in the Solar System, and provides an indirect probe of the internal structure of such planets.

Both the red colour and the H $\alpha$  emission from LkCa 15 b can be explained by the presence of an accreting circumplanetary disk, but some caveats should be kept in mind. Our knowledge of H $\alpha$  emission from accretion disks builds on data from disks around young stars that are hundreds of times more massive than planets. Sallum and co-workers therefore extrapolate the known relationship between H $\alpha$  flux and disk-accretion rate to a completely new size scale. Measurements of other accretion tracers would be desirable, such as 'continuum' emissions at ultraviolet and optical

wavelengths<sup>11</sup>. The nature of the two sources that do not emit H $\alpha$  photons also remains unclear. Follow-up observations, especially at mid-infrared and submillimetre wavelengths, are needed to clear up these issues.

Nevertheless, the authors have demonstrated a powerful technique to find young

planets in circumstellar disks, one that will discover many such planets in the future. This would potentially allow the distribution and occurrence of young planets to be determined with a comparable accuracy to that for the mature exoplanets discovered by the Kepler satellite. Such an understanding of the young planet population will shed light on the decades-old problem of planet formation, and reveal how young planetary systems can evolve into older ones such as our Solar System, billions of years after they were born. ■

Zhaohuan Zhu is in the Department of Astrophysical Sciences, Princeton University, Princeton, New Jersey 08544, USA.  
e-mail: zhzh@astro.princeton.edu

1. Sallum, S. *et al.* *Nature* **527**, 342–344 (2015).
2. Borucki, W. *et al.* *Science* **327**, 977–980 (2010).
3. Espaillat, C. *et al.* *Protostars and Planets VI* 497 (Univ. Arizona Press, 2014).
4. Kraus, A. L. & Ireland, M. J. *Astrophys. J.* **745**, 5 (2012).
5. Quanz, S. P. *et al.* *Astrophys. J.* **766**, L1 (2013).
6. Reggiani, M. *et al.* *Astrophys. J.* **792**, L23 (2014).
7. Zhu, Z. *Astrophys. J.* **799**, 16 (2015).
8. Lovelace, R. V. E., Covey, K. R. & Lloyd, J. P. *Astron. J.* **141**, 51 (2011).
9. Hartmann, L., Hewett, R. & Calvet, N. *Astrophys. J.* **426**, 669 (1994).
10. Close, L. M. *et al.* *Proc. SPIE* **9148**, 91481M (2014).
11. Calvet, N. & Gullbring, E. *Astrophys. J.* **509**, 802 (1998).

## EVOLUTION

# On the crest of becoming vertebrate

The discovery of cells in an invertebrate that share several features with vertebrate neural-crest cells provides insights into how this vital vertebrate cell population might have evolved. [SEE LETTER P.371](#)

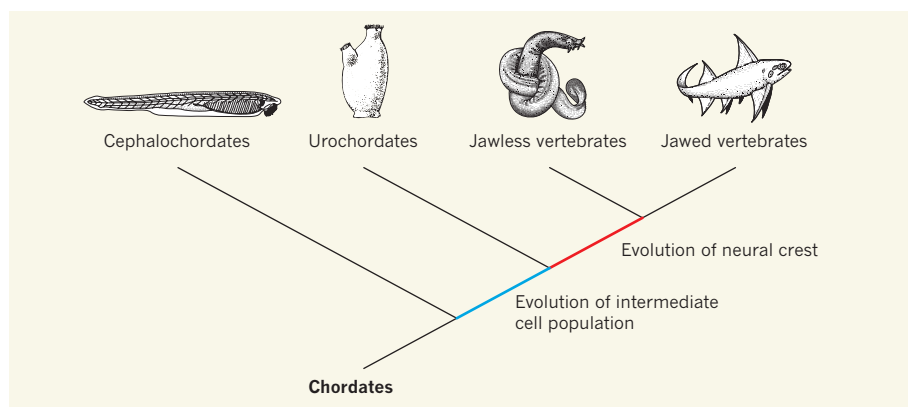
MARIANNE E. BRONNER

The evolution of vertebrates is intimately linked to the advent of an embryonic cell population called the neural crest. Neural-crest cells arise in the central nervous system (CNS) and then invade the periphery of the vertebrate embryo, where they differentiate to form a wide range of cell types, from facial cartilage and bone, to pigment cells of the skin, to neurons and glial cells of the peripheral nervous system. The evolution of these features is thought to have imbued vertebrates with their predatory ability<sup>1</sup>, facilitating their success on Earth. Although all vertebrates have neural-crest cells, how this population evolved has remained a mystery. In this issue, Stolfi *et al.*<sup>2</sup> (page 371) report that a type of neuron in an ascidian (sea squirt) — an invertebrate

filter feeder that is a close relative of vertebrates — has intriguingly similar features to neurons derived from the vertebrate neural crest.

The neural crest is characterized by its origin in the CNS, its migratory behaviour and its ability to differentiate into many cell types (multipotency). Although all animals of the chordate phylum, which includes vertebrates, have a similar body plan (including a dorsal CNS; a structure called the notochord that runs down the midline of the embryo; and a segmented trunk), invertebrates lack cells that have all the characteristics of the neural crest. So far, no intermediate cell type — which would be expected to originate from the CNS and become migratory — has been identified in invertebrates.

Stolfi and colleagues investigated the origin of bipolar tail neurons (BTNs) in the ascidian



**Figure 1 | Evolution of the neural crest.** This simplified phylogenetic tree depicts the evolution of the chordate lineage. An embryonic cell population called the neural crest, which migrates extensively and forms many cell types, arose with the advent of vertebrates. Whereas jawless and jawed vertebrates possess neural-crest cells, this population is not present in cephalochordates or urochordates. Stolfi *et al.*<sup>2</sup> have demonstrated that a cell population in the urochordate *Ciona intestinalis* shares some characteristics with vertebrate neural-crest cells. This may be an intermediate cell population from which the neural crest evolved. (Drawings taken from ref. 8.)

*Ciona intestinalis*. They found that BTNs arise from precursor cells that originate in the developing CNS and migrate through adjacent tissues before differentiating to form mature neurons. Furthermore, the authors discovered that BTNs have a similar function to sensory neurons in vertebrates.

This study is a good complement to previous work<sup>3</sup> demonstrating that precursors to pigment cells are also present in ascidians. These precursors normally remain in the CNS, but can be induced to migrate through misexpression of just one gene, *Twist*. Taken together, these two studies point to the intriguing hypothesis that the evolution of vertebrates might have involved precursor cells in the CNS of a chordate ancestor gaining the ability to form many cell types — having already developed differentiation programs for forming some neural-crest derivatives, including sensory neurons and pigment cells. Moreover, by demonstrating that BTN precursors migrate from the CNS before differentiating, Stolfi *et al.* provide evidence that these cells have analogous traits to two major characteristics of the vertebrate neural crest.

The authors also found that BTN precursors have similar genes to those that encode the vertebrate transcription factors Neurogenin and Islet, which are both required for the formation of sensory neurons. However, the cells do not express all the genes expressed by the neural crest, raising the possibility that BTN precursors represent an intermediate cell population capable of some, but not all, neural-crest-like behaviours.

In contrast to ascidians, more-primitive chordates such as amphioxus, a cephalochordate, lack any precursor cells with neural-crest-like characteristics. These species do, however, have neurons and pigment cells<sup>4</sup>. Moreover, the genomes of all invertebrate chordates, including amphioxus<sup>5</sup>, harbour genes that are similar to those involved in neural-crest formation in vertebrates. Thus, vertebrate

evolution did not require the invention of new genes. Rather, progressive changes that rewired the regulation of gene circuits by using existing factors in new ways probably permitted each transition, from cephalochordates that have no neural crest, to urochordates such as ascidians that have an intermediate cell population with some neural-crest-like characteristics, to the base of vertebrate evolution, with the advent of bona fide neural-crest cells (Fig. 1).

Key questions remain to be answered. For example, it is not clear whether the gene-regulatory networks in the sensory-neuron precursors of ascidians are similar to those of vertebrates. Because signalling pathways that mediate differentiation are thought<sup>6</sup> to be crucial for conferring multipotency on the neural crest, it will be interesting to determine

how these signalling pathways came under the regulatory control of the genes that induce the formation of neural-crest cells in vertebrates. Finally, although *C. intestinalis* is useful for experimental analysis, it is highly derived — its genome is much simpler than that of other chordates. Thus, analogous studies should be performed in other, less-derived ascidians, and in other chordates at different positions on the evolutionary tree.

The idea that cell-differentiation programs have come under the progressive control of neural-crest genes during vertebrate evolution is not new<sup>7</sup>. It has long been thought likely that evolutionary precursors of neural-crest cells lacked the multipotency that defines this cell population in vertebrates. Stolfi and colleagues' study neatly demonstrates the existence of an intermediate cell type that might have gained this ability, enabling the evolution of the neural crest. ■

**Marianne E. Bronner** is in the Division of Biology, California Institute of Technology, Pasadena, California 91125, USA.  
e-mail: mbronner@caltech.edu

1. Gans, C. & Northcutt, R. G. *Science* **220**, 268–273 (1983).
2. Stolfi, A., Ryan, K., Meinertzhagen, I. A. & Christiaen, L. *Nature* **527**, 371–374 (2015).
3. Abitua, P. B., Wagner, E., Navarrete, I. A. & Levine, M. *Nature* **492**, 104–107 (2012).
4. Holland, L. Z. & Holland, N. D. *Curr. Opin. Neurobiol.* **9**, 596–602 (1999).
5. Yu, J.-K., Meulemans, D., McKeown, S. J. & Bronner-Fraser, M. *Genome Res.* **18**, 1127–1132 (2008).
6. Anderson, D. J. *Neuron* **3**, 1–12 (1989).
7. Green, S. A., Simoes-Costa, M. & Bronner, M. E. *Nature* **520**, 474–482 (2015).
8. Sansom, R. S., Gabbott, S. E. & Purnell, M. A. *Nature* **463**, 797–800 (2010).

This article was published online on 28 October 2015.

## CANCER

# Organ-seeking vesicles

**An analysis reveals that cancer cells remotely prepare distant sites for tumour spread in an organ-specific manner, by deploying organ-seeking extracellular vesicles. [SEE ARTICLE P.329](#)**

**JANUSZ RAK**

**T**he metastatic dissemination of cancer cells from their site of origin through the bloodstream to distant organs is a major cause of cancer-related deaths. This process is not random<sup>1</sup>; instead, certain populations of cancer cells preferentially seek out and colonize specific organs<sup>2</sup>, under the control of a range of molecular programs<sup>3</sup>. Such homing implicitly involves interactions between cancer cells that escape the primary tumour, sometimes known as seeds, and the microenvironment, or 'soil', of

target sites<sup>1</sup>. But less intuitive is the discovery by Hoshino *et al.*<sup>4</sup>, described on page 329 of this issue, that seeds can influence the soil before their arrival, sending out extracellular vesicles called exosomes that precondition specific organs for metastatic invasion.

There is growing support for the provocative notion that a build-up of systemic responses to a primary tumour might precede, and even enable, the eruption of metastatic cancer. These responses might involve complex alterations in the body's vascular, coagulation and inflammatory systems — for example,



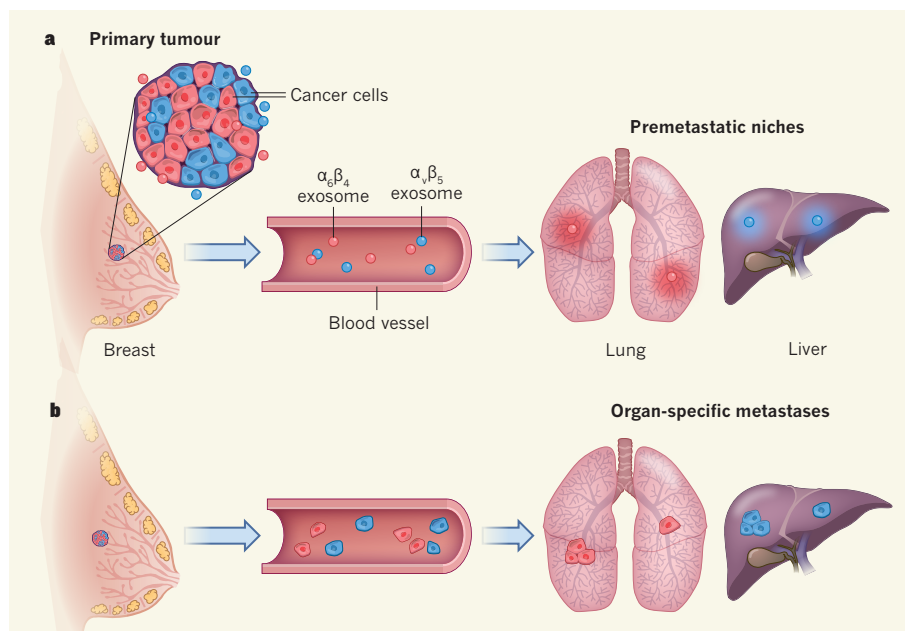
cancer-related changes in the composition of soluble proteins, in cell populations<sup>3</sup> or in the characteristics of exosomes<sup>5</sup> in the blood.

Hoshino *et al.* define exosomes as small extracellular vesicles<sup>6</sup> — membrane-bounded compartments that transport proteins, lipids and nucleic acids<sup>7</sup> from one cell to another, and which can travel considerable distances in bodily fluids or the bloodstream. This information-transfer process has attracted considerable interest in cancer research, because some extracellular vesicles carry cancer-causing genes called oncogenes, or oncogenic proteins that promote cancer formation and disease progression<sup>8</sup>.

The involvement of extracellular vesicles, including exosomes, in metastasis has been studied for some time<sup>9,10</sup>, and contributes to several key events that prepare a distant site for colonization — a process called pre-metastatic niche formation<sup>11</sup>. For example, in a mouse model of melanoma, contact between exosomes and the capillary wall triggers vascular permeability, which enables cancer cells to escape from the blood vessel into a new site<sup>5</sup>. In addition, these exosomes can transfer the oncogenic MET receptor protein to circulating blood cells called myeloid cells, altering the cells' behaviour such that they condition premetastatic sites for subsequent colonization by cancer cells<sup>5</sup>. In pancreatic cancer, circulating exosomes transfer migration inhibitory factor protein to immune cells called Kupffer cells in the liver, triggering a cascade of events that results in premetastatic niche formation<sup>12</sup>.

Although these results indicate that exosomes can promote metastasis in general, whether and how exosomes are involved in organ-specific metastasis has not been extensively investigated. To explore this question, Hoshino *et al.* asked whether cancer-cell types known to preferentially home to the lung, liver, brain or bone might produce exosomes that selectively interact with the same organ. Remarkably, this is precisely what they observed. When exosomes from cancer cells were injected into mice, they became lodged in the organ to which those cancer cells are prone to metastasize. Furthermore, the organ-seeking exosomes interacted with different cell types. For instance, exosomes that targeted the lung became lodged in the epithelial cells that line the organ's interior, whereas liver-targeting exosomes entered Kupffer cells.

Hoshino *et al.* injected mice with exosomes followed by cancer cells from the same cell line, and demonstrated that the exosomes promoted organ-specific metastatic growth. They then made a tantalizing observation — exosomes taken from breast-cancer cells that metastasize to the lung could redirect another cancer-cell population to disseminate in the lung, when it would normally home to the bone. This discovery strengthens the notion that the metastatic characteristics of cancer cells are not autonomous, but can



**Figure 1 | Paving the way for organ-specific metastasis.** **a**, Small extracellular vesicles called exosomes bud off from cancer cells in a primary tumour and enter the bloodstream, transporting proteins, lipids and nucleic acids to distant cells in the body. Hoshino *et al.*<sup>4</sup> report that exosomes derived from different cell types within a mixed population of cancer cells can display different integrin proteins on their surface. This integrin profile promotes adhesion with cells at specific target sites — exosomes displaying the integrin  $\alpha_v\beta_4$  preferentially interact with cells in the lung, whereas  $\alpha_v\beta_3$  directs exosomes to the liver. **b**, The contents of the exosome trigger cellular changes in the target organ that condition the site for metastasis. Thus, exosomes promote organ-specific invasion and metastatic growth of the cancer-cell type from which they originated.

instead be influenced by external factors.

The authors provide several clues to how exosomes orchestrate organ-specific metastasis. They found that exosomes targeting different sites displayed different cell-adhesion receptor proteins called integrins on their surface. The integrin profile of each exosome subtype facilitated its uptake into organs in which an abundance of ligand for that integrin was produced. For instance,  $\alpha_v\beta_3$  integrin directed exosomes to the liver, whereas  $\alpha_v\beta_4$  promoted homing to the lung (Fig. 1). Furthermore, inhibiting the exosomal expression or binding of integrins limited organ-specific metastasis. Finally, the authors found evidence that invasion of target organs by exosomes triggered the production of S100 proteins, which promote inflammation and cell migration, and activation of the protein Src — responses that precondition host cells for metastasis.

These fascinating observations expand our understanding of organ-specific metastasis. However, further investigation is required to establish whether and how this knowledge can be put to practical use. The authors demonstrate that integrin expression might predict metastatic spread, pointing to the possibility that exosomal integrin profiles could be used in cancer diagnostics. Their data also indicate that integrin inhibitors might curtail metastatic spread to specific organs. But in many cases, advanced cancers disseminate to several sites<sup>3</sup>, limiting the potential of therapeutics that work in an organ-specific manner.

It is also worth considering that the molecular pathways that induce metastasis, both exosome-dependent and -independent, are probably extremely diverse. As such, they might be triggered by many context-specific factors: the activation of differentiation pathways in cancer cells; the emergence of a particular molecular subtype within a tumour; therapeutic interventions and more. For example, the incidence of brain metastasis differs between molecular subtypes of breast cancer, and tends to be higher in those driven by the oncogenic protein ERBB2, even after effective treatment with ERBB2 inhibitors<sup>13</sup>. Whether, and how, ERBB2, its antagonists and the therapies used to treat ERBB2-driven cancers might influence the emission of organ-seeking exosomes is unknown, and is of great interest. Similarly, inflammation, abnormal clotting and other cancer-associated changes in physiology might interfere with the organ-seeking mechanism of exosomes and cells — and must be taken into account when analysing routes of metastasis. Thus, much remains to be understood about the fascinating part that organ-specific exosomes might play in fertilizing the metastatic soil in different human cancers. ■

**Janusz Rak** is in the Department of Pediatrics, McGill University, Montreal, Quebec H4A 3J1, Canada, and at the Research Institute of McGill University Health Centre, Montreal Children's Hospital.  
e-mail: janusz.rak@mcgill.ca

1. Paget, S. *Lancet* **1**, 571–573 (1889).
2. Fidler, I. J. *Nature Rev. Cancer* **3**, 453–458 (2003).
3. Nguyen, D. X., Bos, P. D. & Massagué, J. *Nature Rev. Cancer* **9**, 274–284 (2009).
4. Hoshino, A. *et al.* *Nature* **527**, 329–335 (2015).
5. Peinado, H. *et al.* *Nature Med.* **18**, 883–891 (2012).
6. Lötval, J. *et al.* *J. Extracell. Vesicles* **3**, 26913 (2014).
7. Colombo, M., Raposo, G. & Théry, C. *Annu. Rev. Cell Dev. Biol.* **30**, 255–289 (2014).
8. Rak, J. *Front. Pharmacol.* **4**, 21 (2013).
9. Hood, J. L., San, R. S. & Wickline, S. A. *Cancer Res.* **71**, 3792–3801 (2011).
10. Poste, G. & Nicolson, G. L. *Proc. Natl Acad. Sci. USA* **77**, 399–403 (1980).

11. Kaplan, R. N. *et al.* *Nature* **438**, 820–827 (2005).
12. Costa-Silva, B. *et al.* *Nature Cell Biol.* **17**, 816–826 (2015).
13. Steeg, P. S., Camphausen, K. A. & Smith, Q. R. *Nature Rev. Cancer* **11**, 352–363 (2011).

This article was published online on 28 October 2015.

## REHABILITATION

# Boost for movement

**By electrically stimulating the motor neurons of rats that have spinal-cord injury, in bursts that are attuned to the times at which the neurons receive voluntary motor commands, the animals' recovery can be improved.**

RANDOLPH J. NUDDO

People with spinal-cord injury face a host of challenges, including sensory, bowel, bladder and sexual dysfunction, paralysis and weakness. Although rehabilitation can help to improve motor and sensory function, recovery is limited. But writing in *Proceedings of the National Academy of Sciences*, McPherson *et al.*<sup>1</sup> show that rats with spinal-cord injury can recover substantial motor ability when treated with a new type of electrical-stimulation therapy.

Rehabilitation from spinal-cord injury relies on the damaged nervous system adapting over time. For example, when a person with motor defects in their arm repeatedly practises hand grasps, their ability to do this improves as a result of changes in neural circuitry. This alteration in nervous communication patterns and connections to regain function is called neuronal plasticity. Therapeutic approaches that encourage neuronal plasticity therefore have the potential to improve recovery when combined with more-conventional rehabilitation strategies.

In one such approach, known as electrical stimulation, a barrage of electrical excitation is given to undamaged spinal-cord projections (called fibres) and the neurons that they target, boosting the body's own weak electrical inputs to these circuits. In addition to promoting neuronal plasticity, this technique can help to uncover latent motor functions that are presumably inactive owing to insufficient levels of excitation. For example, it has been used to demonstrate<sup>2</sup> that even people with 'complete' spinal-cord injuries (who are paralysed below the level of the injury) retain some latent functional potential.

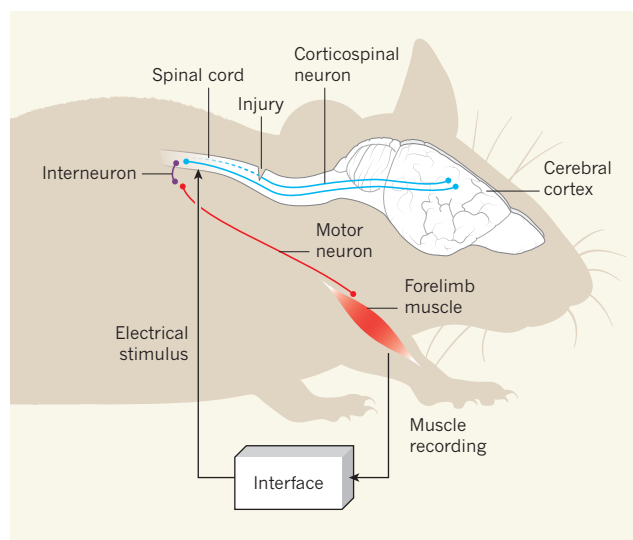
In the mid-twentieth century, the neuro-psychologist Donald Hebb predicted that the synaptic connections between neurons can be altered, becoming stronger or weaker depending on the timing of their activation<sup>3</sup>. If the presynaptic neuron is activated before the postsynaptic neuron, then the synaptic connection is strengthened, increasing activation of the postsynaptic neuron in response to signals from the presynaptic neuron. This theory is summarized in a modern maxim<sup>4</sup>: "cells that fire together, wire together." Thanks to improvements in microelectronics, it is now

possible to create small devices that record electrical biological signals and then use those signals to trigger rapid electrical stimulation in the nervous system, allowing Hebb's theory to be tested directly in living organisms. A pivotal 2006 experiment<sup>5</sup> in monkeys demonstrated that artificial coupling of areas in the brain's cerebral cortex can alter the function of the connected neurons. More-recent studies<sup>6–9</sup> have begun to explore whether timed stimulation can strengthen synaptic connections in the brain and spinal cord.

McPherson and colleagues experimented with a variation on this technique in injured rats. They recorded an electrical signal from muscle that indicated that the rat was in the process of contracting that muscle, and used the signal to trigger an electrical-stimulation pulse in the spinal cord (Fig. 1). They hypothesized that, in this way, voluntary commands to induce muscle contraction (which are transmitted from the cerebral cortex and perhaps from other brain centres) would arrive at motor neurons in the spinal cord shortly before the neurons were activated by the stimulation pulse. This would strengthen the synaptic connections in the spinal cord. Once the connections had been strengthened, the voluntary commands from the brain might be sufficient to drive muscle contraction on their own.

First, the authors injured one side of the spinal cord of rats. This resulted in paralysis of the forelimb on the injured side, followed by limited natural recovery. The researchers then inserted microwires into the spinal cord, below the level of the injury. Such microwires can stimulate neurons, and even movement, with extremely low electrical currents (approximately ten times lower than that needed for neuronal stimulation at the surface of the spinal cord), allowing specific activation of motor neurons<sup>10</sup>. Wires were also implanted in forelimb muscles to allow recording of electrical signals and were used to trigger spinal-cord stimulation.

Six weeks after spinal-cord injury, the rats were unable to reach and grasp a small food pellet with the affected forelimb. All the animals were severely impaired, achieving only about 13% of their pre-injury scores in this reaching task. Over



**Figure 1 | Precisely timed spinal-cord stimulation.** Voluntary-movement commands are conveyed to the spinal cord by corticospinal neurons that originate in the motor regions of the brain's cerebral cortex. These neurons terminate on spinal-cord interneurons, which in turn project to motor neurons that excite muscle. After spinal-cord injury, corticospinal neurons are damaged, impairing limb function. However, some neurons can escape injury. McPherson *et al.*<sup>1</sup> investigated recovery from spinal-cord injury in rats. An integrated electronic interface recorded electrical activity in limb muscle (indicating contraction of the muscles) and applied a carefully timed electrical stimulus to the spinal cord. The stimulus arrived shortly after the weakened voluntary command from the surviving corticospinal neurons. This excitation helped to restore limb motor function, possibly by strengthening the neuronal connections in the spinal cord. (Adapted from ref. 1.)



the following 13 weeks, they underwent rehabilitation training, practising reaching for 30 minutes each day, 5 days per week. One group of rats received only rehabilitative training; a second group received rehabilitation and electrical stimulation in the spinal cord that was triggered by muscle activity; a third group received rehabilitation and electrical stimulation in the spinal cord that was not linked to the timing of muscle activity.

Rats in the second group demonstrated substantial recovery. At the end of the regime, they had regained 63% of their pre-injury motor ability. By contrast, rats in the first and third control groups made much smaller improvements. Perhaps most importantly, the second group maintained motor function after a three-week follow-up period. By timing spinal-cord stimulation to coincide with the 'sweet spot' for motor-neuron activation, synaptic connections seemed to have been strengthened over time, allowing enhanced motor recovery.

Although spinal-cord circuitry is complex, and the exact mechanisms that underlie this dramatic effect remain to be fleshed out, McPherson and colleagues' results have substantial clinical relevance. This study is an example of therapeutic strategies for neural repair that go beyond simply exciting the injured nervous system with a blast of electrical stimulation. By mirroring the timing rules that the nervous system normally uses to enhance the strength of synaptic connections, it might be possible to guide neuronal plasticity in a

more functional and adaptive way after injury.

Nevertheless, many theoretical and practical issues need to be addressed before such approaches are attempted in humans. For example, the safety of chronic electrical stimulation using electrodes in the spinal cord must be thoroughly tested. The timing of the stimulation may also need to be adjusted on the basis of differences in nerve-fibre size and speed of conduction in the human nervous system compared with that of rats. Nonetheless, the translation of techniques that exploit synaptic plasticity from bench to bedside is now a little closer. ■

**Randolph J. Nudo** is in the Department of Rehabilitation Medicine, University of Kansas Medical Center, Landon Center on Aging, Kansas City, Kansas 66160, USA.  
e-mail: rnudo@kumc.edu

1. McPherson, J. G., Miller, R. R. & Perlmutter, S. I. *Proc. Natl Acad. Sci. USA* **112**, 12193–12198 (2015).
2. Angeli, C. A., Edgerton, V. R., Gerasimenko, Y. P. & Harkema, S. J. *Brain* **137**, 1394–1409 (2014).
3. Hebb, D. O. *The Organization of Behavior* (Wiley, 1949).
4. Shatz, C. J. *Sci. Am.* **267**, 60–67 (1992).
5. Jackson, A., Mavoori, J. & Fetzi, E. E. *Nature* **444**, 56–60 (2006).
6. Nishimura, Y., Perlmutter, S. I., Eaton, R. W. & Fetzi, E. E. *Neuron* **80**, 1301–1309 (2013).
7. Rebesch, J. M. & Miller, L. E. *Prog. Brain Res.* **192**, 83–102 (2011).
8. Lucas, T. H. & Fetzi, E. E. *J. Neurosci.* **33**, 5261–5274 (2013).
9. Guggenmos, D. J. et al. *Proc. Natl Acad. Sci. USA* **110**, 21177–21182 (2013).
10. Mushahwar, V. K. & Horch, K. W. *Ann. NY Acad. Sci.* **860**, 531–535 (1998).

## ECOLOGY

# Ecosystem responses to climate extremes

**Extreme drought or wet conditions have now been found to strongly influence the vegetative development of ecosystems. Semi-arid regions are most affected — raising concerns about their vulnerability to long-term drought in the future.**

ANJA RAMMIG & MIGUEL D. MAHECHA

**E**xtrême climatic conditions such as drought or heatwaves are likely to intensify in the next few decades<sup>1</sup>. Long-term observations<sup>2</sup> of past decades suggest that characteristic recurrence frequencies, intensities and durations of certain extreme events have already increased noticeably. One pressing question is whether key ecosystem services, such as the capacity of ecosystems to accumulate carbon, are affected by extreme events<sup>3</sup>. The potential of ecosystems to accumulate carbon is intimately related to the phenology of vegetation<sup>4</sup> — essentially, all the characteristic

periodicities in an organism's life, such as the annual cycles of plant leaf development. Writing in the *Journal of Geophysical Research*, Ma et al.<sup>5</sup> describe how climate extremes modify the seasonal vegetation development of different ecosystems.

Ma and colleagues' study relies mainly on a measure called the enhanced vegetation index. This is used to determine whether a region of Earth contains live green vegetation, and can be derived from spectral data that have been collected by satellites for more than a decade. Vegetation indices of this kind are often used as indicators of vegetation productivity, although they cannot really be directly



## 50 Years Ago

In the last of nine papers presented to ... the British Association for the Advancement of Science ... Dr. K. Adam discussed the responsibility of television: this responsibility has since been sharpened by the continuing discussions on the feasibility and the scope for a 'university of the air' ... it is proposed that only the major conurbations, containing 60–70 per cent of the population, should be covered by broadcast television ... coverage in this way does not mean that the public beyond the range of television and sound transmission would be unable to enrol at the University of the Air ... all the programmes provided for television by the University should be recorded and prints of the original transmission can readily be circulated anywhere for viewing ... All the courses would be supported by carefully prepared lecture notes and reading lists, and each week all enrolled students would be required to submit exercises for marking and comment by tutors.

From *Nature* 20 November 1965

## 100 Years Ago

The little attention given to science in education and in the public mind has been the theme of many essays and addresses ... science is usually regarded as suitable for study by a select few only, and not as an essential part of all modern life and thought ... We do not for a moment suggest that the end of all education should be preparation for scientific careers; neither do we ask that men of letters, statesmen, and administrative officers of departments of State should all be scientific experts ... Our claim is that everyone — from elementary-school pupil to college don — should be made acquainted with appropriate outlines of scientific work and thought.

From *Nature* 18 November 1915



**Figure 1 | Hummock grassland during dry and wet growth seasons.** Some ecosystems, such as hummock grasslands in southeastern Australia, compensate for poor growth during dry periods by increasing growth during wet periods. Ma *et al.*<sup>5</sup> report that such ecosystems have until now been less vulnerable to drought than are croplands or pasture.

interpreted as such. Nevertheless, such indices reveal the annual cycle of vegetation-canopy development within and across ecosystems.

Drought stress is also commonly described by an index that allows an intuitive interpretation of an ecosystem's water balance. In combination with temperature data, indices for vegetation and drought allow researchers to describe the complex interplay between vegetation and climate, as Ma and colleagues do in their work.

In contrast to previous studies, which mainly examined the phenology of individual ecosystems on subcontinental scales, Ma *et al.* revealed the impacts of climatic extremes on the phenology of different ecosystem types on a continental scale. They show that two extreme years, 2002 (extremely dry) and 2010 (extremely wet), had significant effects on vegetation phenology and productivity across southeastern Australia — a region characterized by large rainfall and temperature gradients, and which has vegetation types ranging from arid grasslands to forests. The authors find that 70% of the area had a prolonged growing season during 2010, whereas there was essentially no observable growing season in 2002. These observations indicate the fundamental relevance of fluctuations in water availability for phenology.

A crucial question emerges from this work: which vegetation types and ecosystems are most vulnerable to extreme conditions? By analysing the change in the vegetation index relative to changes in drought conditions for each location, Ma and co-workers suggest that cropland and pastures are most sensitive to fluctuations in climatic conditions. Natural systems, such as the hummock grasslands investigated in the study, adapt to high year-to-year variations in rainfall by increasing growth during wet periods to compensate for poor growth during dry periods (Fig. 1). But agricultural systems cannot do this — they

respond only to dry periods by reducing growth, which supports the view that agricultural systems might be strongly affected by extreme climatic events.

The authors show that drought sensitivity peaks in semi-arid regions across different climate regimes. This is of interest, because semi-arid ecosystems cover approximately 40% of the global land surface<sup>6</sup>. Ecosystems of this kind are often dominated by grasses or shrubs that have developed mechanisms to survive long periods of dryness, and which grow rapidly when rainfall arrives. Ma *et al.* show that growth was almost completely dormant during 2002, and increased swiftly when the rains arrived during 2010. A remaining question is how long this behaviour can be maintained during severe droughts that last for more than one season. In other words, how resilient are ecosystems to increasing fluctuations in climate? With additional pressure from human land-use activities, the resilience of many of these ecosystems may be on a knife-edge<sup>7</sup>.

Ma and co-workers note that their findings have serious implications for estimates of the effects of extreme weather events on the global carbon cycle. But whether and how these results can be extrapolated to the global scale is unclear. Semi-arid regions have recently gained attention<sup>8,9</sup> for their ability to take up large amounts of carbon dioxide from the atmosphere during favourable conditions, thereby driving the variability in the rate of increase in CO<sub>2</sub> levels in the atmosphere. But it is not known whether, or for how long, semi-arid regions and other terrestrial carbon sinks will continue to absorb so much CO<sub>2</sub>, nor what the magnitude of this carbon sink is — particularly given the uncertainty in predicting changes in the frequency and intensity of extreme events<sup>3</sup>.

Overall, the authors' results demonstrate the need for a better understanding of plant responses to increased climate variability if

regional and global models are to be improved. Obtaining such an understanding will require a multifaceted approach. For instance, differences in the size of responses might be partly attributable to varying rates of biodiversity<sup>10</sup>, a factor that cannot be considered on the basis of the relatively coarse spectral information contained in conventional vegetation indices. And although Ma and colleagues' study provides much-needed insight into ecosystem behaviour in response to extreme events, we can still only speculate about the exact nature of the underlying ecophysiological mechanisms and the potential legacy of such events. ■

Anja Rammig is at the TUM School of Life Sciences Weihenstephan, Technische Universität München, 85354 Freising, Germany. Miguel D. Mahecha is at the Max Planck Institute for Biogeochemistry, 07745 Jena, Germany.  
e-mails: anja.rammig@tum.de; mmahecha@bgc-jena.mpg.de

1. Seneviratne, S. I. *et al.* in *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change* (eds Field, C. B. *et al.*) 109–230 (Cambridge Univ. Press, 2012).
2. Hartmann, D. L. *et al.* in *Climate Change 2013: The Physical Science Basis. Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Stocker, T. F. *et al.*) 159–254 (Cambridge Univ. Press, 2013).
3. Reichstein, M. *et al.* *Nature* **500**, 287–295 (2013).
4. Richardson, A. D. *et al.* *Agric. For. Meteorol.* **169**, 156–173 (2013).
5. Ma, X., Huete, A., Moran, S., Ponce-Campos, G. & Eamus, D. J. *Geophys. Res. Biogeosci.* <http://dx.doi.org/10.1002/2015JG003144> (2015).
6. UN Environment Management Group. *Global Drylands: A UN System-Wide Response* (UN, 2011).
7. Holmgren, M. *et al.* *Front. Ecol. Environ.* **4**, 87–95 (2006).
8. Ahlström, A. *et al.* *Science* **348**, 895–899 (2015).
9. Poulter, B. *et al.* *Nature* **509**, 600–603 (2014).
10. Isbell, F. *et al.* *Nature* **526**, 574–577 (2015).



# Dpp spreading is required for medial but not for lateral wing disc growth

Stefan Harmansa<sup>1</sup>, Fisun Hamaratoglu<sup>2</sup>, Markus Affolter<sup>1\*</sup> & Emmanuel Caussinus<sup>1,3\*</sup>

***Drosophila* Decapentaplegic (Dpp) has served as a paradigm to study morphogen-dependent growth control. However, the role of a Dpp gradient in tissue growth remains highly controversial. Two fundamentally different models have been proposed: the ‘temporal rule’ model suggests that all cells of the wing imaginal disc divide upon a 50% increase in Dpp signalling, whereas the ‘growth equalization model’ suggests that Dpp is only essential for proliferation control of the central cells. Here, to discriminate between these two models, we generated and used morphotrap, a membrane-tethered anti-green fluorescent protein (GFP) nanobody, which enables immobilization of enhanced (e)GFP::Dpp on the cell surface, thereby abolishing Dpp gradient formation. We find that in the absence of Dpp spreading, wing disc patterning is lost; however, lateral cells still divide at normal rates. These data are consistent with the growth equalization model, but do not fit a global temporal rule model in the wing imaginal disc.**

Morphogens regulate patterning and growth of tissues and organs by forming long-range gradients from regions of high concentration (the source) to regions of low concentration (the adjacent target field)<sup>1–5</sup>. In *Drosophila*, the vertebrate bone morphogenetic protein (BMP)2/4 homologue Dpp is studied extensively in the wing imaginal disc. This larval precursor of the fly wing and the dorsal thorax is subdivided into an anterior and a posterior compartment<sup>6,7</sup>. Dpp is expressed in a stripe of anterior cells adjacent to the compartment boundary<sup>8</sup>, forming long-range anterior and posterior extracellular gradients in the target field<sup>9–11</sup>. The Dpp gradient is transduced by its receptors Thickveins (Tkv)<sup>12</sup> and Punt<sup>13</sup> and translated into an intracellular gradient of phosphorylated Mothers against dpp (p-Mad)<sup>14</sup>. Dpp signalling suppresses transcription of *brinker* (*brk*)<sup>15–17</sup>, a repressor of Dpp target gene transcription<sup>14</sup> and a repressor of growth<sup>18</sup>. This results in high p-Mad levels (high Dpp signalling) in the medial region of the wing disc and high Brk levels (low Dpp signalling) in the lateral region of the wing disc. The interplay of p-Mad and Brk coordinates the expression profiles of other Dpp targets, such as *spalt* (*sal*), *optomotor blind* (*omb*; also known as *bifid*) and *daughters against dpp* (*dad*)<sup>19–21</sup>. In addition to its role in patterning, Dpp is a key regulator of growth; overexpression of Dpp promotes wing disc overgrowth<sup>22,23</sup>, while *dpp* mutant wing discs remain very small<sup>24</sup>.

To our knowledge, the requirement for Dpp spreading has never been explicitly tested by, for example, blocking Dpp dispersal by tethering it to the cell membrane, as has been done for the Wingless (Wg) morphogen<sup>25–27</sup>. The available experimental evidence strongly supports an instructive and essential role for Dpp spreading in the control of patterning (reviewed in refs 14, 28, 29). However, the role of Dpp spreading in growth control is highly controversial<sup>15,28,30,31</sup>. Two major models have been suggested to explain how the Dpp gradient controls uniform proliferation and growth of the wing disc. One model, the temporal rule, suggests that all cells of the wing imaginal disc compute the level of Dpp and divide upon a 50% increase in Dpp signalling. In contrast, the growth equalization model proposes that Dpp sustains the proliferation of medial cells by the removal of the growth repressor Brk, while the proliferation rate of lateral cells is limited by Brk to rates that can be sustained by medial cells, resulting

in a uniform proliferation profile along the wing disc tissue<sup>28,32–34</sup>. In the growth equalization model, the Dpp/Brk system is not a growth promoter but is rather a growth-modulatory system, ironing out inherent regional differences in proliferation rates<sup>32</sup>. To study the role of Dpp spreading in wing disc patterning and growth better, we designed and experimentally established a novel approach to manipulate morphogen spreading *in vivo*.

## Nanobody-mediated morphogen trapping

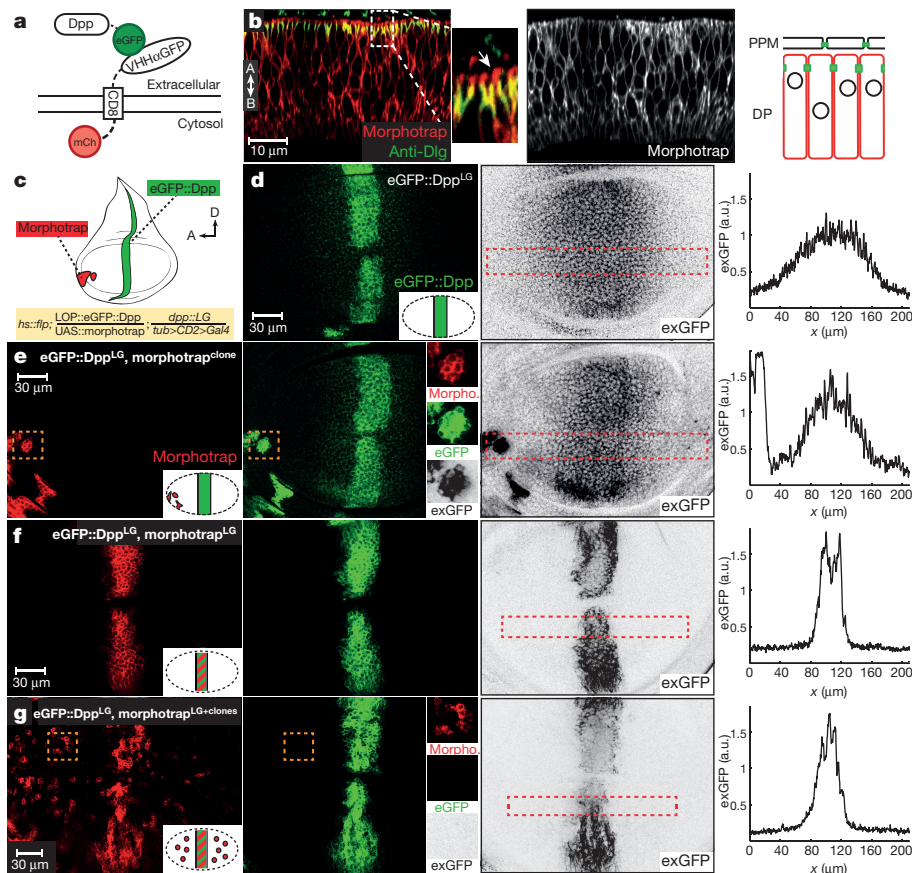
To manipulate the Dpp gradient *in vivo*, we designed and implemented a synthetic morphogen trapping system consisting of a GFP-tagged morphogen (in our case, eGFP::Dpp) and a generic extracellular GFP trap (VHH-GFP4::CD8::mCherry; referred to as morphotrap) (Fig. 1a). Our eGFP::Dpp construct is based on a previously published fusion protein<sup>9</sup> and was implemented as a LexA inducible transgene (see Methods and Extended Data Fig. 1a). Morphotrap (VHH-GFP4::CD8::mCherry) represents a fusion protein consisting of an extracellular, single-domain nanobody against GFP<sup>35</sup> (and cognate fluorescent tags, including eGFP), followed by the mouse CD8 transmembrane domain and a cytoplasmic mCherry fluorescent tag. Morphotrap was implemented as a Gal4-inducible transgene as well as a LexA-inducible transgene (see Methods). The principle idea behind morphotrap is to immobilize the extracellular fraction of eGFP::Dpp in *Drosophila* tissues in a controlled spatial manner, either in the presence or the absence of wild-type Dpp (Fig. 1a).

Expression of eGFP::Dpp by the *dpp-LG* LexA driver line in a *dpp*<sup>d8/d12</sup> mutant background restored proper Dpp signalling in the wing tissue such that the size and pattern was rescued to a large extent and adult flies developed (Extended Data Fig. 1). These results show that our eGFP::Dpp fusion protein acts as a good surrogate for Dpp in the wing disc.

Morphotrap localizes along the basolateral and apical surface of wing disc cells (Fig. 1b), and does not interfere with Dpp signalling or cell survival when expressed at high levels (Extended Data Fig. 2a–e). Therefore, morphotrap can be expressed at high levels and accumulates around the expressing wing disc cells without interfering with cell division and patterning.

<sup>1</sup>Growth & Development, Biozentrum, Klingelbergstrasse 50/70, University of Basel, 4056 Basel, Switzerland. <sup>2</sup>Center for Integrative Genomics, University of Lausanne, 1015 Lausanne, Switzerland. <sup>3</sup>Institute of Molecular Life Sciences (IMLS), University of Zurich, 8057 Zurich, Switzerland.

\*These authors contributed equally to this work.



**Figure 1 | Morphotrap can block eGFP::Dpp spreading.** **a**, The morphotrap system. **b**, Optical cross-section (left, middle) and schematics (right) along the apical-basal (A-B) axis of a wing disc expressing morphotrap in disc proper (DP) cells only (*nubbin::Gal4*). Morphotrap is localized all along the cell membrane (in red), both apical (arrow) and basal to the junctional marker Discs-large (Dlg; in green). peripodial membrane (PPM). **c**, Schematic representation of eGFP::Dpp (LexA/LOP) and morphotrap clones (*Gal4/UAS*). For all discs anterior (A) is oriented to the left and dorsal (D) is oriented to the top. **d**, A wild-type

wing disc expressing eGFP::Dpp in the Dpp stripe (*dpp::LG*), visualized by eGFP fluorescence (left) or by extracellular GFP staining (exGFP) (middle). Fluorescence intensity profile of the region marked by a red rectangle (right). a.u., arbitrary units. **e**, Lateral morphotrap clones trap extracellular eGFP::Dpp. **f**, Gradient formation is blocked by co-expression of eGFP::Dpp and morphotrap in the Dpp stripe (both expressed by *dpp::LG*). **g**, Co-expression of eGFP::Dpp and morphotrap in the stripe fully blocks Dpp spreading since additional morphotrap clones do not show eGFP signal (see insets in second panel from the left).

## Morphotrap can modify the Dpp gradient

We then tested whether exposing morphotrap on the cell surface locally modified the extracellular concentration of eGFP::Dpp. We generated random small clones of morphotrap in wild-type wing discs expressing eGFP::Dpp in the domain of *dpp* transcription. To set apart the induced morphotrap clones from the cells expressing eGFP::Dpp, we used *Gal4* and *LexA* drivers to induce morphotrap and eGFP::Dpp, respectively (Fig. 1c; see Methods). In control discs, in which no clones were generated, eGFP::Dpp formed a bilateral extracellular concentration gradient visualized by sensitive extracellular immunostainings against eGFP (Fig. 1d). The eGFP signal dropped below detection levels at a distance of approximately 60  $\mu$ m from the medial eGFP::Dpp source. In discs in which small clones expressing morphotrap had been generated, we detected high levels of extracellular eGFP::Dpp coating the surface of the clone cells, even when the clones were located in regions in which eGFP::Dpp was not detected otherwise (Fig. 1e). These results show that morphotrap is able to sequester extracellular eGFP::Dpp, even in areas of low or non-detectable eGFP::Dpp.

Trapped eGFP::Dpp was active in signalling, since morphotrap clones located in the lateral region of the disc showed increased p-Mad levels, mainly along the edge facing the eGFP::Dpp source (Extended Data Fig. 2f, g). The results show that eGFP::Dpp disperses over the entire width of the disc, although its levels cannot normally be detected above background levels in the lateral regions using fluorescent microscopy (see also ref. 30). We conclude that eGFP::Dpp can

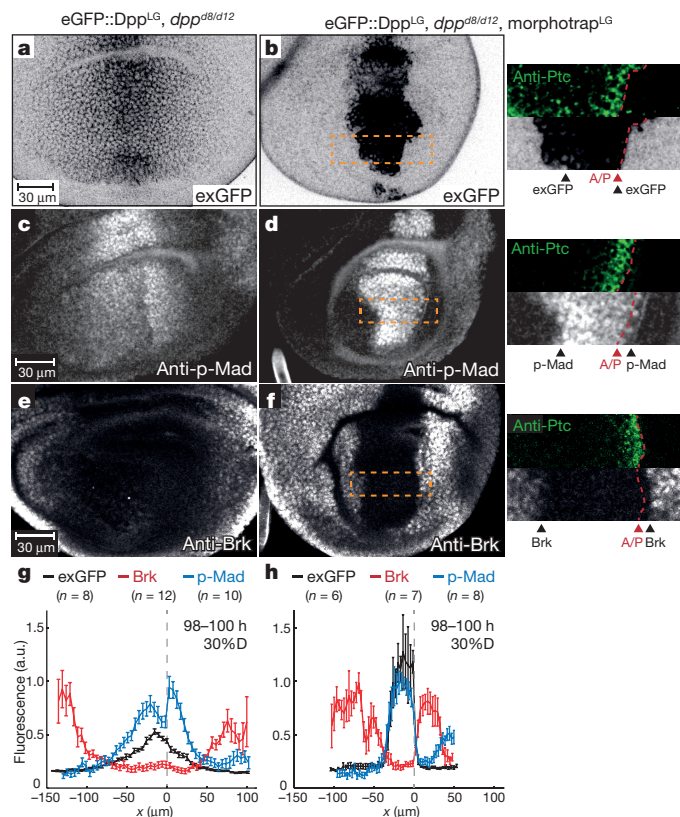
interact with its receptors when bound to the cell surface by morphotrap and that lateral cells can respond to Dpp.

To investigate whether morphotrap was able to interfere with the formation of the extracellular concentration gradient of eGFP::Dpp when expressed in the source cells, we expressed both eGFP::Dpp and morphotrap in wild-type wing discs in the domain of *dpp* transcription. Under these conditions, we did not detect any dispersal of eGFP::Dpp using antibody staining (Fig. 1f), suggesting that eGFP::Dpp cannot leave the source region owing to tethering to secreting cells. Furthermore, clones expressing morphotrap in lateral cells did not accumulate any eGFP::Dpp on the cell surface in these conditions, neither did clones in the vicinity of the eGFP::Dpp source (Fig. 1g, middle; see insets). These results demonstrate that morphotrap fully retains eGFP::Dpp on source cells and completely abolishes the formation of the extracellular concentration gradient of eGFP::Dpp.

## Dpp spreading is required for patterning

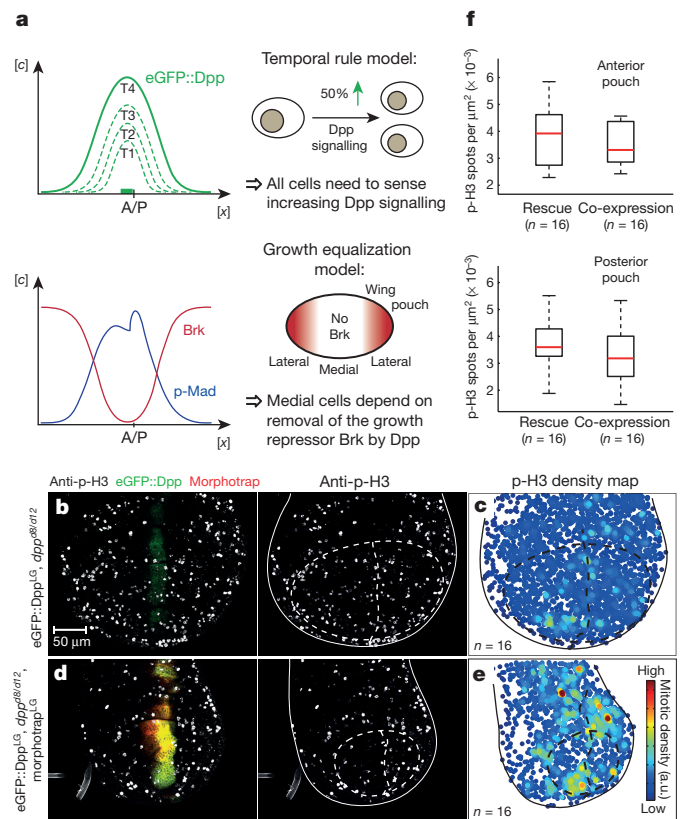
The function of *dpp* for patterning the wing disc has been studied extensively<sup>11,36,37</sup>. However, how a loss of Dpp spreading would affect target gene expression has not been tested directly. To compare Dpp signalling responses in control *dpp*<sup>d8/d12</sup> wing discs rescued by eGFP::Dpp (eGFP::Dpp gradient is present) to Dpp signalling responses in *dpp*<sup>d8/d12</sup> wing discs expressing both eGFP::Dpp and morphotrap in the expression domain of *dpp* (eGFP::Dpp gradient





**Figure 2 | Blocking Dpp spreading results in a sharp p-Mad/Brk transition.** **a**, Representative *dpp<sup>d8/d12</sup>* mutant wing disc rescued with eGFP::Dpp (rescue) and stained for exGFP (grey). **b**, Left, exGFP signal in a disc co-expressing eGFP::Dpp and morphotrap (*dpp-LG*, co-expression) in a *dpp<sup>d8/d12</sup>* mutant. Right, magnification of the region marked by the rectangle on the left, showing that all signal is from anterior cells where Dpp is expressed. A/P boundary is determined by Ptc staining (green) and marked by a dotted line. Approximate domain size is marked by arrowheads. **c**, **d**, p-Mad staining in rescue (**c**) and co-expression (**d**) wing discs. **e**, **f**, Brk staining in rescue (**e**) and co-expression (**f**) wing discs. **g**–**f**, Average fluorescence intensity profiles of 98–100 h after egg laying (AEL) old larvae measured to the edge of the wing disc of rescued (**g**) and co-expression (**h**) wing discs. Profiles were measured with 30% dorsal (D) offset parallel to the dorso/ventral (D/V) boundary (see Methods for details). Error bars show standard deviation (s.d.).

is absent; Methods and Fig. 2), we performed immunostainings against p-Mad, Brk, Sal and Omb. In control discs, p-Mad, Sal and Omb formed three bilateral gradients of different widths, Sal being the narrowest and Omb being the widest (Fig. 2c, g and Extended Data Fig. 3a); Brk was only detected in the most lateral regions of the discs (Fig. 2e, g). In contrast, when eGFP::Dpp and morphotrap were co-expressed, Dpp spreading and hence gradient formation was fully blocked throughout development (Extended Data Fig. 4a–c). In these discs, the p-Mad, Sal and Omb gradients collapsed in the posterior compartment to a single row of cells abutting the anterior source of eGFP::Dpp (Fig. 2d, h and Extended Data Fig. 3b); high levels of Brk were detected in the posterior compartment up to the source of Dpp, except for a single row of cells abutting the compartment boundary (Fig. 2f, h). Similar results were obtained regarding target gene expression in the anterior compartment upon trapping eGFP::Dpp in source cells (Fig. 2h and Extended Data Fig. 3). In addition, we inhibited eGFP::Dpp dispersal in posterior cells only (Extended Data Fig. 5); under these conditions p-Mad failed to form a posterior long-range gradient and both Sal and Omb expression collapsed onto the narrow p-Mad domain. Hence, wing disc patterning in the posterior compartment was abolished (Extended Data Fig. 5a–f). Wings of flies with blocked or reduced Dpp spreading lacked proper



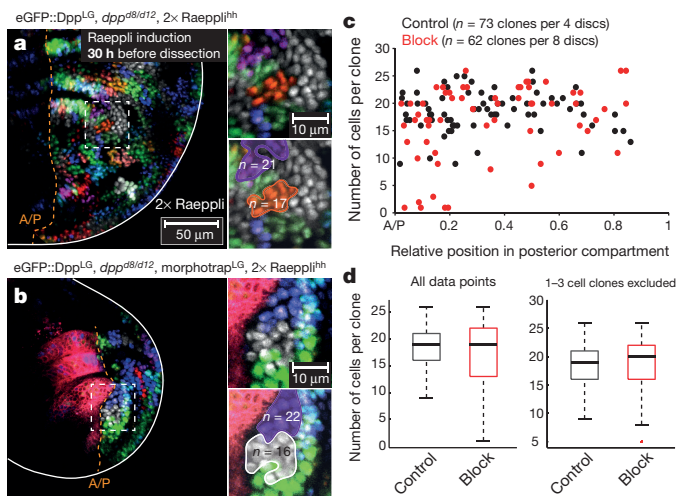
**Figure 3 | The uniform proliferation pattern is independent of Dpp spreading.** **a**, Top, the temporal rule model of growth control. Bottom, the growth equalization model. **b**, p-H3 staining in a representative *dpp<sup>d8/d12</sup>* mutant wing disc rescued with eGFP::Dpp. The A/P boundary and the pouch outline are marked by dotted lines (right). **c**, **e**, Computed p-H3 spots density (of  $n = 16$  discs) in rescue (**c**) and co-expression (**e**) wing discs (see Methods). **d**, p-H3 signal in a *dpp<sup>d8/d12</sup>* mutant wing disc co-expressing eGFP::Dpp and morphotrap. **f**, Mitotic density in the anterior ( $P > 0.05$ ) and posterior pouch ( $P > 0.05$ ); whiskers correspond to minimum and maximum data points.

wing vein patterning (Extended Data Figs 3f and 9d, f). Altogether, these results show that dispersal of Dpp is strictly required for the patterning function of Dpp.

### Dpp spreading and growth control

Despite numerous studies addressing the role of Dpp in the control of growth of the wing imaginal disc, the conclusions drawn from different sets of experiments have resulted in conflicting interpretations. In the temporal rule model<sup>30,38</sup>, all disc cells compute the increase in Dpp levels and divide upon a gain of 50%. In sharp contrast, the growth equalization model<sup>28</sup> proposes that lateral cells proliferate independently of Dpp (Fig. 3a). In line with this later model, Dpp signalling has been blocked in regions outside of the wing pouch in several studies, without much effect on cell proliferation<sup>39,40</sup>. However, it has not been possible to directly modulate the Dpp gradient at the protein level until now, making it difficult to interpret the requirement of Dpp long-range function in growth control.

To discriminate between these two growth control models, we aimed at using a different experimental approach, directly eliminating the Dpp gradient at the protein level using morphotrap. As described earlier, the elimination of the gradient leads to the absence of Dpp signalling, that is, the target genes *sal* and *omb* are not expressed in the wing epithelium beyond the source cells and the immediate neighbours, and the Brk repressor is present at high levels in all cells beyond the Dpp source. We thus compared the proliferation pattern of control *dpp<sup>d8/d12</sup>* wing discs rescued by eGFP::Dpp to the proliferation pattern of *dpp<sup>d8/d12</sup>* wing discs expressing both eGFP::Dpp and morphotrap in the

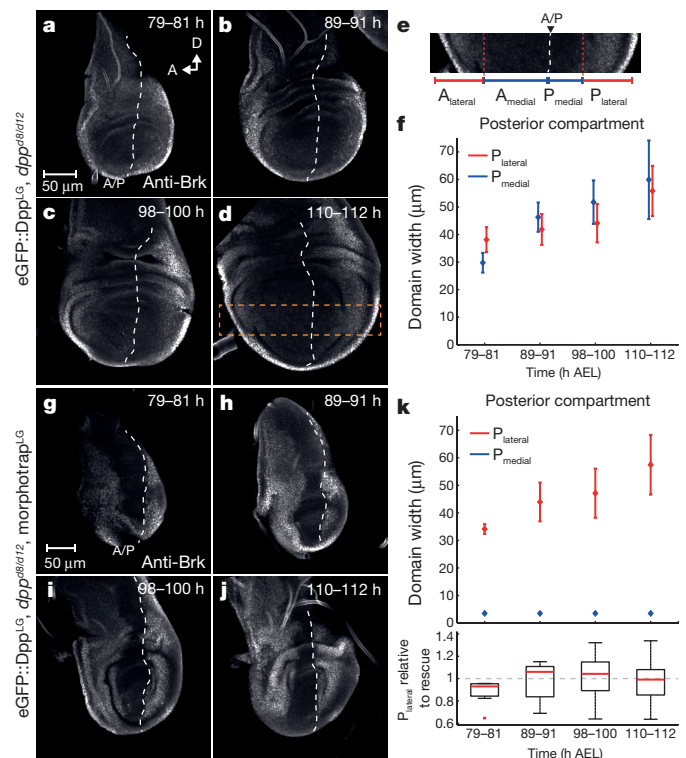


**Figure 4 | Block of Dpp spreading does not affect clonal proliferation rates.** **a–d**, To estimate clonal proliferation rates in the posterior compartment, Raeppli was induced 30 h before dissection (66–70 h AEL) and larvae were dissected at 96–100 h AEL. **a**, Control disc. **b**, Disc with blocked Dpp spreading. **c**, Cell numbers per clone were counted and plotted against the relative position in the posterior compartment (0 corresponding to the A/P boundary and 1 to the posterior edge of the disc). **d**, Boxplots showing the number of cells per clone for all data points (left plot,  $P > 0.05$ ) or when the 1–3 cell clones are excluded (right plot,  $P > 0.05$ ).

expression domain of *dpp*, that is, we compared the growth rates of wing disc cells in the presence and in the absence of eGFP::Dpp spreading. We visualized the proliferation pattern of such wing discs by staining for phospho-histone H3 (p-H3), a marker for mitotic cells. In wild-type wing discs, cell proliferation was shown to be rather homogeneous in third instar wing discs<sup>7,41,42</sup>. Our quantitative analyses showed that in discs rescued with eGFP::Dpp, the proliferation profile was also uniform (Fig. 3b, c). Interestingly, blocking Dpp spreading neither affected the uniform proliferation pattern (Fig. 3d, e) nor did we detect significant changes in the mitotic density in wing imaginal discs during the observed developmental stages (Fig. 3f and Extended Data Fig. 4g–i).

To obtain a more global and more quantitative view of the cell division patterns, we used the whole-tissue labelling tool Raeppli<sup>43</sup> to induce differently marked clones in control wing discs and in wing discs in which Dpp spreading was blocked by morphotrap (Fig. 4 and Extended Data Fig. 6). To compare the proliferation rates in the presence or absence of Dpp spreading, we induced colour selection in clones at different time points of development and quantitatively evaluated the resulting clone size after defined time points (number of cells per clone). In control wing discs, clonal growth rates were homogeneous along the anterior–posterior (A/P) axis (Fig. 4c, black dots). When Dpp spreading was blocked, we observed that the majority of clones showed similar growth rates to control clones, and we did not find a significant difference in clonal proliferation between controls and discs with blocked Dpp spreading (Fig. 4d). However, with blocked Dpp spreading, we also found low numbers of small clones (1–3 cells) next to the A/P boundary (Extended Data Fig. 7). These small clones were not found in control discs, in which Dpp spreading was normal. The presence of such small clones might hint towards the fact that a subpopulation of wing disc cells depend on Dpp signalling to divide and/or survive.

Both the p-H3 data and the Raeppli results demonstrate that the cells in the lateral Brk domain do not depend on Dpp spreading to proliferate (in contradiction with the temporal rule model), but rather that the proliferation rate is set by a Dpp-independent system (consistent with the growth equalization model).



**Figure 5 | The development of the medial but not the lateral wing disc requires Dpp spreading.** **a–k**, Data set from 79–112 h AEL stained for Brk. **a–d**, Representative rescued wing discs of four time points investigated. **e**, Magnification of area marked in **d**, visualizing the location of medial (high Dpp signalling) and lateral (low Dpp signalling) domain (see Methods). **f**, Temporal development of domain width in the posterior compartment in rescued discs. **g–j**, Representative co-expression wing discs. **k**, Temporal development of domain width in co-expression wing discs, and size change of the lateral domain relative to control discs (rescue  $n = 34$ , co-expression  $n = 37$ ). **f**, **k**, Error bars show s.d.

## Dpp spreading and size control

Using morphotrap in *dpp* mutant flies also allowed us to address how long-range spreading of Dpp affects wing disc size control. We quantified and compared the temporal growth profile of the posterior compartment of control *dpp<sup>d8/d12</sup>* wing discs rescued by eGFP::Dpp to the growth profile of *dpp<sup>d8/d12</sup>* wing discs co-expressing both eGFP::Dpp and morphotrap in the expression domain of *dpp*. We performed immunostainings against Brk at different time points between 80 and 112 h after egg laying (AEL). In control discs, the posterior compartment doubled in width during the observed time window (Fig. 5a–d and Extended Data Fig. 8d). We delimited a medial low Brk (indicating high Dpp signalling) zone and a lateral high Brk (indicating low Dpp signalling) zone (Fig. 5e; see Methods); both zones increased in width at the same speed, keeping a constant relative proportion of 1:1 (Fig. 5f), consistent with published data<sup>44</sup>. In discs in which spreading of Dpp was abolished, the low Brk zone in the centre of the disc was reduced to a single medial row of cells in the posterior compartment (see earlier and Fig. 5g–j). During the observed time window, the lateral part of the posterior compartment showed similar widths and width increases as the lateral high Brk zone of the posterior compartment of control discs (Fig. 5k). Similar growth profiles were seen in discs expressing eGFP::Dpp in the source stripe and morphotrap in the posterior compartment (Extended Data Fig. 5g). These results demonstrate that growth in the lateral region of the wing disc is independent of the extracellular Dpp gradient and does not depend on the dynamics of Dpp signalling. In support of this finding, similar growth dynamics were observed for the anterior compartment (Extended Data Fig. 8a, b).



In contrast, the medial, Brk-negative region is lost when Dpp spreading is blocked, suggesting that Dpp dispersal is important for growth control of the medial region, in particular in the central wing pouch area. We therefore quantified wing pouch size using the inner Wg-expression ring as a pouch marker. We measured the size of the pouch in *dpp* mutant discs rescued with eGFP::Dpp, and compared it to the pouch of discs in which either eGFP::Dpp dispersal was hindered in the posterior compartment only, or in which the release of eGFP::Dpp from the anterior source was completely blocked (Extended Data Fig. 9). Upon hindering Dpp spreading in the posterior compartment, the size of the posterior pouch was reduced by approximately 40%. Strikingly, when we trapped eGFP::Dpp in the source, the size of the posterior wing pouch was even further reduced (by more than 60%). These results indicate that eGFP::Dpp spreading is essential for wing pouch growth. The analyses using the whole-tissue labelling technique Raeppli (Extended Data Fig. 7) further showed that small clones were found in the posterior compartment close to the compartment boundary when morphotrap is expressed in source cells. Such clones were not found in control discs. Together, these data show that Dpp signalling has an important role in proliferation control of medial wing pouch cells, as indicated by earlier studies<sup>33,39,40</sup>, and further suggest that the range of Dpp spreading might be crucially linked to the size of the wing pouch region along the A/P axis.

## Discussion

We used morphotrap, a novel approach to manipulate the extracellular Dpp gradient in the wing imaginal disc. Expressing morphotrap in small clones of lateral wing disc cells captures eGFP::Dpp in regions of the disc in which eGFP::Dpp cannot be detected above background levels. This finding demonstrates that Dpp does disperse over the entire wing imaginal disc, and that low Dpp levels could control cell behaviour even in lateral regions. However, we find that while Dpp spreading is strictly required for wing disc patterning, it is not essential for cell proliferation in the lateral region of the wing disc. These results are consistent with the growth equalization model but are in disagreement with a disc-wide temporal rule model, and suggest that lateral cells do not compute Dpp signalling levels to trigger cell division. It has been argued that Dpp-independent Dpp signalling (in addition to Dpp-dependent Dpp signalling) might control cell proliferation according to the temporal rule model<sup>45</sup>. This interpretation was based on the observation that in genetic experiments in which Dpp signalling was eliminated by the concomitant genetic removal of *brk* and *tkv* (or *brk* and *dpp*), certain Dpp targets were active owing to the absence of the potent Brk repressor<sup>31,38</sup>. However, in our experiments using morphotrap, Dpp signalling was eliminated via the removal of the Dpp gradient and led to the absence of Dpp target gene expression and to the presence of high levels of Brk in the entire lateral wing disc. Therefore, in our experimental setting, Dpp signalling was turned off in the lateral cells, yet these cells divided at a normal rate, as quantitatively shown by our experiments using Raeppli. As cell division should be abolished (or altered) in the absence of Dpp signalling, according to the temporal rule, our experiments reject a general, disc-wide temporal rule model for wing disc growth control.

However, our data are entirely consistent with the proposal of the growth equalization model, suggesting that Dpp spreading results in medial removal of Brk and that this repression of *brk* represents an essential step in the formation of the wing pouch tissue<sup>33</sup>. Our results support the growth equalization model that the wing disc tissue consists of two regions with different requirements for Dpp signalling, namely a medial region that depends on Dpp signalling to grow and a lateral region that grows independent of Dpp.

While the growth equalization model does not explain final organ size, our results suggest that the range of Dpp spreading is linked to the size of the wing pouch (albeit not to the entire disc). In a number of elegant studies, the range of Wg signalling was suggested to control

pouch growth via a feed-forward recruitment mechanism<sup>27,46</sup>, presumably together with Dpp. Interestingly, the replacement of the major endogenous *Drosophila* Wnt, Wg, with one that expresses a membrane-tethered form of the protein, showed that Wg spreading and gradient formation is dispensable for patterning and to some extent for growth of the pouch<sup>26</sup>. In contrast, our results on Dpp strongly support the notion that Dpp spreading is essential for its role in pouch patterning and size control. Getting a better understanding of the control of wing pouch growth will require the combinatorial manipulation of the Dpp and the Wg signalling pathways to study individual pathway outputs as well as their mutual interactions at different time points throughout larval development. Furthermore, it will be of major importance to study the interactions of the morphogen systems with other growth control systems (for example, the insulin–phosphatidylinositol-3-OH kinase and the Hippo pathways) to better understand the control of final organ size<sup>47</sup>. The addition of the morphotrap and the Raeppli techniques to such analyses will help gain better insight into how morphogens control organ growth.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 1 March; accepted 10 September 2015.**

**Published online 9 November 2015.**

- Ashe, H. L. & Briscoe, J. The interpretation of morphogen gradients. *Development* **133**, 385–394 (2006).
- Baena-Lopez, L. A., Nojima, H. & Vincent, J. P. Integration of morphogen signalling within the growth regulatory network. *Curr. Opin. Cell Biol.* **24**, 166–172 (2012).
- Rogers, K. W. & Schier, A. F. Morphogen gradients: from generation to interpretation. *Annu. Rev. Cell Dev. Biol.* **27**, 377–407 (2011).
- Schwank, G. & Basler, K. Regulation of organ growth by morphogen gradients. *Cold Spring Harb. Perspect. Biol.* **2**, a001669 (2010).
- Wartlick, O., Mumcu, P., Jülicher, F. & Gonzalez-Gaitan, M. Understanding morphogenetic growth control—lessons from flies. *Nature Rev. Mol. Cell Biol.* **12**, 594–604 (2011).
- Martin, F. A., Herrera, S. C. & Morata, G. Cell competition, growth and size control in the *Drosophila* wing imaginal disc. *Development* **136**, 3747–3756 (2009).
- Garcia-Bellido, A. & Merriam, J. R. Parameters of the wing imaginal disc development of *Drosophila melanogaster*. *Dev. Biol.* **24**, 61–87 (1971).
- Masucci, J. D., Miltenberger, R. J. & Hoffmann, F. M. Pattern-specific expression of the *Drosophila* decapentaplegic gene in imaginal disks is regulated by 3' cis-regulatory elements. *Genes Dev.* **4**, 2011–2023 (1990).
- Teleman, A. A. & Cohen, S. M. Dpp gradient formation in the *Drosophila* wing imaginal disc. *Cell* **103**, 971–980 (2000).
- Entchev, E. V., Schwabedissen, A. & González-Gaitán, M. Gradient formation of the TGF- $\beta$  homolog Dpp. *Cell* **103**, 981–992 (2000).
- Nellen, D., Burke, R., Struhl, G. & Basler, K. Direct and long-range action of a Dpp morphogen gradient. *Cell* **85**, 357–368 (1996).
- Nellen, D., Affolter, M. & Basler, K. Receptor serine/threonine kinases implicated in the control of *Drosophila* body pattern by decapentaplegic. *Cell* **78**, 225–237 (1994).
- Ruberte, E., Marty, T., Nellen, D., Affolter, M. & Basler, K. An absolute requirement for both the type II and type I receptors, punt and thick veins, for Dpp signaling *in vivo*. *Cell* **80**, 889–897 (1995).
- Affolter, M. & Basler, K. The Decapentaplegic morphogen gradient: from pattern formation to growth regulation. *Nature Rev. Genet.* **8**, 663–674 (2007).
- Jaźwińska, A., Kirov, N., Wieschaus, E., Roth, S. & Rushlow, C. The *Drosophila* gene *brinker* reveals a novel mechanism of Dpp target gene regulation. *Cell* **96**, 563–573 (1999).
- Minami, M., Kinoshita, N., Kamoshida, Y., Tanimoto, H. & Tabata, T. *brinker* is a target of Dpp in *Drosophila* that negatively regulates Dpp-dependent genes. *Nature* **398**, 242–246 (1999).
- Campbell, G. & Tomlinson, A. Transducing the Dpp morphogen gradient in the wing of *Drosophila*: regulation of Dpp targets by *brinker*. *Cell* **96**, 553–562 (1999).
- Doumpas, N. *et al.* Brk regulates wing disc growth in part via repression of Myc expression. *EMBO Rep.* **14**, 261–268 (2013).
- Barrio, R. & de Celis, J. F. Regulation of spalt expression in the *Drosophila* wing blade in response to the Decapentaplegic signaling pathway. *Proc. Natl Acad. Sci. USA* **101**, 6021–6026 (2004).
- Weiss, A. *et al.* A conserved activation element in BMP signaling during *Drosophila* development. *Nature Struct. Mol. Biol.* **17**, 69–76 (2010).
- Sivasankaran, R., Vigano, M. A., Müller, B., Affolter, M. & Basler, K. Direct transcriptional control of the Dpp target *omb* by the DNA binding protein Brinker. *EMBO J.* **19**, 6162–6172 (2000).

22. Capdevila, J. & Guerrero, I. Targeted expression of the signaling molecule decapentaplegic induces pattern duplications and growth alterations in *Drosophila* wings. *EMBO J.* **13**, 4459–4468 (1994).
23. Zecca, M., Basler, K. & Struhl, G. Sequential organizing activities of engrailed, hedgehog and decapentaplegic in the *Drosophila* wing. *Development* **121**, 2265–2278 (1995).
24. Spencer, F. A., Hoffmann, F. M. & Gelbart, W. M. Decapentaplegic: a gene complex affecting morphogenesis in *Drosophila melanogaster*. *Cell* **28**, 451–461 (1982).
25. Zecca, M. & Struhl, G. Recruitment of cells into the *Drosophila* wing primordium by a feed-forward circuit of vestigial autoregulation. *Development* **134**, 3001–3010 (2007).
26. Alexandre, C., Baena-Lopez, A. & Vincent, J. P. Patterning and growth control by membrane-tethered Wingless. *Nature* **505**, 180–185 (2014).
27. Zecca, M. & Struhl, G. A feed-forward circuit linking Wingless, Fat-Dachsous signaling, and the Warts-Hippo pathway to *Drosophila* wing growth. *PLoS Biol.* **8**, e1000386 (2010).
28. Restrepo, S., Zartman, J. J. & Basler, K. Coordination of patterning and growth by the morphogen DPP. *Curr. Biol.* **24**, R245–R255 (2014).
29. Hamaratoglu, F., Affolter, M. & Pyrowolakis, G. Dpp/BMP signaling in flies: from molecules to biology. *Semin. Cell Dev. Biol.* **32**, 128–136 (2014).
30. Wartlick, O. *et al.* Dynamics of Dpp signaling and proliferation control. *Science* **331**, 1154–1159 (2011).
31. Schwank, G., Yang, S. F., Restrepo, S. & Basler, K. Comment on “Dynamics of Dpp signaling and proliferation control”. *Science* **335**, 401 (2012).
32. Schwank, G., Restrepo, S. & Basler, K. Growth regulation by Dpp: an essential role for Brinker and a non-essential role for graded signaling levels. *Development* **135**, 4003–4013 (2008).
33. Martín, F. A., Pérez-Garijo, A., Moreno, E. & Morata, G. The *brinker* gradient controls wing growth in *Drosophila*. *Development* **131**, 4921–4930 (2004).
34. Schwank, G. *et al.* Antagonistic growth regulation by Dpp and Fat drives uniform cell proliferation. *Dev. Cell* **20**, 123–130 (2011).
35. Saerens, D. *et al.* Identification of a universal VHH framework to graft non-canonical antigen-binding loops of camel single-domain antibodies. *J. Mol. Biol.* **352**, 597–607 (2005).
36. Lecuit, T. *et al.* Two distinct mechanisms for long-range patterning by Decapentaplegic in the *Drosophila* wing. *Nature* **381**, 387–393 (1996).
37. Müller, B., Hartmann, B., Pyrowolakis, G., Affolter, M. & Basler, K. Conversion of an extracellular Dpp/BMP morphogen gradient into an inverse transcriptional gradient. *Cell* **113**, 221–233 (2003).
38. Wartlick, O., Mumcu, P., Jülicher, F. & Gonzalez-Gaitan, M. Response to Comment on “Dynamics of Dpp Signaling and Proliferation Control”. *Science* **335**, 401 (2012).
39. Martín-Castellanos, C. & Edgar, B. A. A characterization of the effects of Dpp signaling on cell growth and proliferation in the *Drosophila* wing. *Development* **129**, 1003–1013 (2002).
40. Burke, R. & Basler, K. Dpp receptors are autonomously required for cell proliferation in the entire developing *Drosophila* wing. *Development* **122**, 2261–2269 (1996).
41. Milán, M., Campuzano, S. & García-Bellido, A. Cell cycling and patterned cell proliferation in the wing primordium of *Drosophila*. *Proc. Natl Acad. Sci. USA* **93**, 640–645 (1996).
42. Mao, Y. *et al.* Differential proliferation rates generate patterns of mechanical tension that orient tissue growth. *EMBO J.* **32**, 2790–2803 (2013).
43. Kanca, O., Caussinus, E., Denes, A. S., Percival-Smith, A. & Affolter, M. Raeppli: a whole-tissue labeling tool for live imaging of *Drosophila* development. *Development* **141**, 472–480 (2014).
44. Hamaratoglu, F., de Lachapelle, A. M., Pyrowolakis, G., Bergmann, S. & Affolter, M. Dpp signaling activity requires Pentagone to scale with tissue size in the growing *Drosophila* wing imaginal disc. *PLoS Biol.* **9**, e1001182 (2011).
45. Wartlick, O., Jülicher, F. & Gonzalez-Gaitan, M. Growth control by a moving morphogen gradient during *Drosophila* eye development. *Development* **141**, 1884–1893 (2014).
46. Zecca, M. & Struhl, G. Control of *Drosophila* wing growth by the vestigial quadrant enhancer. *Development* **134**, 3011–3020 (2007).
47. Hariharan, I. K. Organ size control: lessons from *Drosophila*. *Dev. Cell* **34**, 255–265 (2015).

**Acknowledgements** We would like to acknowledge the work of the late William (Bill) Gelbart, who initiated the work on Dpp. We thank S. Matsuda, I. Alborelli and H. Belting for discussions; T. Schaffter for help and support with WingJ; the Biozentrum Imaging Core Facility for maintenance of microscopes and support. We are grateful to G. Struhl, K. Basler and G. Pyrowolakis for their input and discussion on the project. We thank K. Basler, S. Cohen, G. Morata, R. Bario and E. Laufer for flies and reagents. S.H. was supported by the ‘Fellowships for Excellence’ International PhD Program in Molecular Life Sciences of the Biozentrum, University of Basel. Funding is also acknowledged from the SystemsX.ch initiative within the framework of the WingX (E.C. and F.H.) and the MorphogenetiX projects (E.C.). F.H. is now supported by a Swiss National Science Foundation (SNSF) Professorship grant (PP00P3\_150682). The work in the laboratory was supported by grants from Cantons Basel-Stadt and Basel-Land, from the SNSF and from SystemsX.ch (M.A.).

**Author Contributions** S.H., E.C., F.H. and M.A. conceived and designed the study. S.H. performed the experiments. S.H. analysed the data. S.H., E.C. and M.A. wrote the paper.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.A. (Markus.Affolter@unibas.ch).



## METHODS

**Fly strains.** The following fly lines were used:  $y^1w^{1118}$  (wild type), *dpp-LG86Fb* and *LOP::mCherry-CAAX* (K. Basler<sup>48</sup>), *tub > CD2,Stop > Gal4* (F. Pignoni). *P[Cre]1b* was obtained from Bloomington. *hh-Gal4*, *dpp-Gal4*, *nub-Gal4*, *dpp<sup>d8</sup>* and *dpp<sup>d12</sup>* are described in FlyBase (<http://www.flybase.org>).

**Genotypes by figure.** Figure 1b: *w; nub-Gal4/UAS-morphotrap*; Fig. 1d: *w; LOP-eGFP::Dpp/+; dpp-LG/+*; Fig. 1e: *yw, hsFlp; tub > CD2,Stop > Gal4, LOP-eGFP::Dpp/UAS-morphotrap; dpp-LG/+*; Fig. 1f: *w; LOP-eGFP::Dpp/LOP-morphotrap; dpp-LG/+*; Fig. 1g: *w hsFlp; LOP-eGFP::Dpp, tub > CD2,Stop > Gal4/LOP/UAS-morphotrap; dpp-LG/+*. Fig. 2a, c, e: *w; LOP-eGFP::Dpp, dpp<sup>d12</sup>/dpp<sup>d8</sup>; dpp-LG/+*; Fig. 2b, d, f: *w; LOP-eGFP::Dpp, dpp<sup>d12</sup>/LOP-morphotrap, dpp<sup>d8</sup>; dpp-LG/+*; Fig. 3b, c: *w; LOP-eGFP::Dpp, dpp<sup>d12</sup>/dpp<sup>d8</sup>; dpp-LG/+*; Fig. 3d, e: *w; LOP-eGFP::Dpp, dpp<sup>d12</sup>/LOP-morphotrap, dpp<sup>d8</sup>; dpp-LG/+*; Fig. 4a: *yw, hsFlp; LOP-eGFP::Dpp, dpp<sup>d12</sup>/dpp<sup>d8</sup>; dpp-LG, hh-Gal4/2 × LOP/UAS::Raeppli*; Fig. 4b: *yw, hsFlp; LOP-eGFP::Dpp, dpp<sup>d12</sup>/LOP-morphotrap, dpp<sup>d8</sup>; dpp-LG, hh-Gal4/2 × LOP/UAS::Raeppli*; Fig. 5a–d: *w; LOP-eGFP::Dpp, dpp<sup>d12</sup>/dpp<sup>d8</sup>; dpp-LG/+*; Fig. 5g–j: *w; LOP-eGFP::Dpp, dpp<sup>d12</sup>/LOP-morphotrap, dpp<sup>d8</sup>; dpp-LG/+*.

**Molecular cloning.** For pUASTLOTTattB\_eGFP::Dpp, GFP was replaced by eGFP in the Dpp-GFP plasmid<sup>9</sup> (obtained from S. Cohen). Then, eGFP::Dpp was inserted in the multiple cloning site of pUASTLOTTattB vector<sup>43</sup> by standard cloning procedures.

For pUASTLOTTattB\_VHH-GFP4::CD8::mCherry, we inserted the VHH-GFP4 fragment after the signal peptide sequence of the mouse CD8 domain in the pUAS::CD8::GFP plasmid<sup>49</sup>. We replaced the GFP by a mCherry (Clontech) and finally cloned the VHH-GFP4::CD8::mCherry fragment into the pUASTLOTTattB vector<sup>43</sup>.

Transgenes were inserted by phiC31-integrase-mediated recombination into the 35B region on the 2nd chromosome. Resulting fly lines are responsive to LexA (ref. 48) and Gal4 (ref. 50) transcriptional activators. By crossing these flies to *Cre<sup>+</sup>*-expressing flies, either the UAS or the LOP site is being excised in a mutually exclusive manner. Excision was screened for by PCR as described previously<sup>43</sup>.

**Creation of wing disc data sets.** Flies were kept in standard fly vials (containing polenta and yeast) in a 26°C incubator. Larvae were staged as described previously<sup>44</sup>. In our data sets, we only included male larvae, which were positively selected for the presence of the genital disc. All male larvae of a collection were dissected and further processed to obtain maximum sample numbers.

**Statistics and data representation.** The phenotypes observed and quantified (pattern and size) differ strikingly from controls; therefore no sample size estimation was performed. However, sample number was chosen to ensure statistical significance, which was assessed using a two-sided Student's *t*-test with unequal variance. No randomization was done, however all larvae of an experiment were kept in the same incubator, as well as dissected and processed together using identical solutions in order to minimize variation between the different experimental groups. Blinding was not possible due to the obvious phenotypes observed. For quantitative measurements, the centre values represent the arithmetic mean and the error bars show standard deviation, except for boxplots (Fig. 3f, Fig. 4d, and Fig. 5k, bottom), where centre value correspond to the median and the whiskers mark the maximum and minimum data points.

**Immunostainings and image acquisition.** Staged larvae were dissected and transferred directly to cold fixative (4% PFA in PBS) and fixed for 20 min at room temperature or 40 min at 4°C (for p-Mad and Brk stainings) rotating. After fixation, discs were extensively washed with PBT (PBS plus 0.3% Triton-X) and blocked in PBTN (PBT plus 2% normal donkey serum; Jackson Immuno Research Laboratories) for 30 min at room temperature, followed by incubation with primary antibody overnight at 4°C. The next day discs were washed in PBT six times for 20 min and incubated in secondary antibody for 1.5 h at room temperature on a rotor. After another round of washes with PBT, samples were mounted in Vectashield (H-1000, Vector Laboratories). All discs of one data set were mounted on the same slide using larval brains as spacers. For all quantitative data sets we made sure that imaging conditions allowed acquisition of data in the linear range (Extended Data Fig. 10). For high-resolution imaging along the z-axis (Fig. 1b), discs were mounted with double-sided tape as spacers to avoid squeezing of the discs. The extracellular GFP staining was done as described previously<sup>51</sup>. Images were acquired on a Leica SP5 confocal microscope (section thickness 1 µm for data sets, 0.13 µm for optical cross-section in Fig. 1b).

**BrdU labelling.** Discs were dissected in Schneider's insect medium, followed by a 1 h incubation in Schneider's plus 75 µg ml<sup>-1</sup> BrdU (Sigma, B5002) at room temperature. This was followed by two 5 min washes in Schneider's and one 5 min wash in PBS. Then discs were fixed for 15 min in PBS plus 4% PFA, followed by another 15 min fixation in PBS plus 4% PFA plus 0.6% Triton-X-100. Discs were permeabilized for 60 min in PBS plus 0.3% Triton-X-100 and transferred to a 1:1 mixture of PBS plus 0.6% Triton-X-100: 4 N HCl for 30 min. This was followed by

extensive washes in PBS plus 0.3% Triton-X-100. Discs were incubated overnight in anti-BrdU (1:100, Becton Dickinson, 347580) in PBS plus 0.3% Triton-X-100. Washing, incubation in secondary antibody and mounting were done as described earlier.

**Antibodies.** rb-anti-p-Mad (1:1,500; E. Laufer<sup>52,53</sup>); rb-anti-phospho-Smad1/5 (1:200; Cell Signaling, 9516S; used in Extended Data Fig. 4d–f); gp-anti-Brk (1:1,000; Gines Morata); rb-anti-Sal (1:40; R. Schuh<sup>54</sup>); rat-anti-Sal (1:700; R. Barrio<sup>55</sup>); rb-anti-Omb (1:1,200; G. O. Plugfelder<sup>56</sup>); m-anti-Wg (also known as 4D4-s; 1:120; DSHB, University of Iowa); m-anti-Ptc (also known as Apa1-s; 1:40; DSHB, University of Iowa); rb-anti-GFP (1:200 for extracellular staining; Abcam ab6556); anti-BrdU (1:100; Becton Dickinson, 347580). All secondary antibodies from the AlexaFluor series were used at 1:750 dilutions except for Alexa405-anti-rb and Alexa680-anti-m, which were used at 1:500 dilutions; CF405S-anti-gp was used 1:1,000 (Sigma-Aldrich).

**Image processing.** Images were processed using ImageJ (National Institutes of Health) software. Concentration profiles in Fig. 1, Extended Data Fig. 3 and Extended Data Fig. 5f, right, were created using the Plot Profile function in ImageJ. Optical cross-section in Fig. 1b was created using the section function in Imaris (Bitplane) software. We made use of the Wg/Ptc co-staining<sup>44</sup>, which outlines the wing pouch (Wg), the D/V boundary (Wg) and the A/P boundary (Ptc, see also Extended Data Fig. 2). Quantification of wing pouch size and extraction of average gradient profiles (Figs 2g, h, 3f and Extended Data Figs 1f, 2d, e, 4c, f, i, 5e, 8f, 9g) were done using the WingJ software<sup>57</sup> (<http://tschaffter.ch/projects/wingj/>). For measuring gradient profiles in WingJ, we used average projections of ten consecutive slices spanning the disc proper epithelium only. Gradient profiles were extracted using WingJ software either only in the pouch (Extended Data Fig. 2) or up to the edge of the wing disc (Fig. 2 and Extended Data Fig. 4), which allowed a better representation of lateral Brk profiles. Profiles were measured with a Sigma of 4px and either 15% ventral offset (for Extended Data Fig. 2e) or 30% dorsal offset (for all other profiles) parallel to the D/V border (marked by the Wg staining). Plotting of average concentration profiles was done applying the Matlab toolbox included in WingJ using the Matlab (Mathworks) software.

**Generation of mitotic density maps.** Wing discs were staged and stained for Wg/Ptc and p-H3, a marker labelling mitotic cells. p-H3-positive nuclei were detected using the Imaris software (Bitplane) spot detection tool; peripodial nuclei were excluded from the following computation. Each disc was marked at 15 landmarks (see Extended Data Fig. 8e). Sixteen discs of one time point were fitted to a reference disc using these landmarks by an affine transformation (least square, Fiji–Landmark correspondence plug-in). All data points of these 16 discs were included in a scatter plot using the Scatplot script (A. Sanchez-Barba; <http://www.mathworks.com/matlabcentral/fileexchange/8577-scatplot>) in Matlab. The Scatplot visualizes data point density by a colour map, with high-density regions appearing in red and low-density regions in blue.

The mitotic density in Fig. 3f was calculated by normalizing the number of p-H3-positive cells in the anterior or posterior pouch to the corresponding pouch area. Statistical significance was assessed using a two-sided Student's *t*-test with unequal variance.

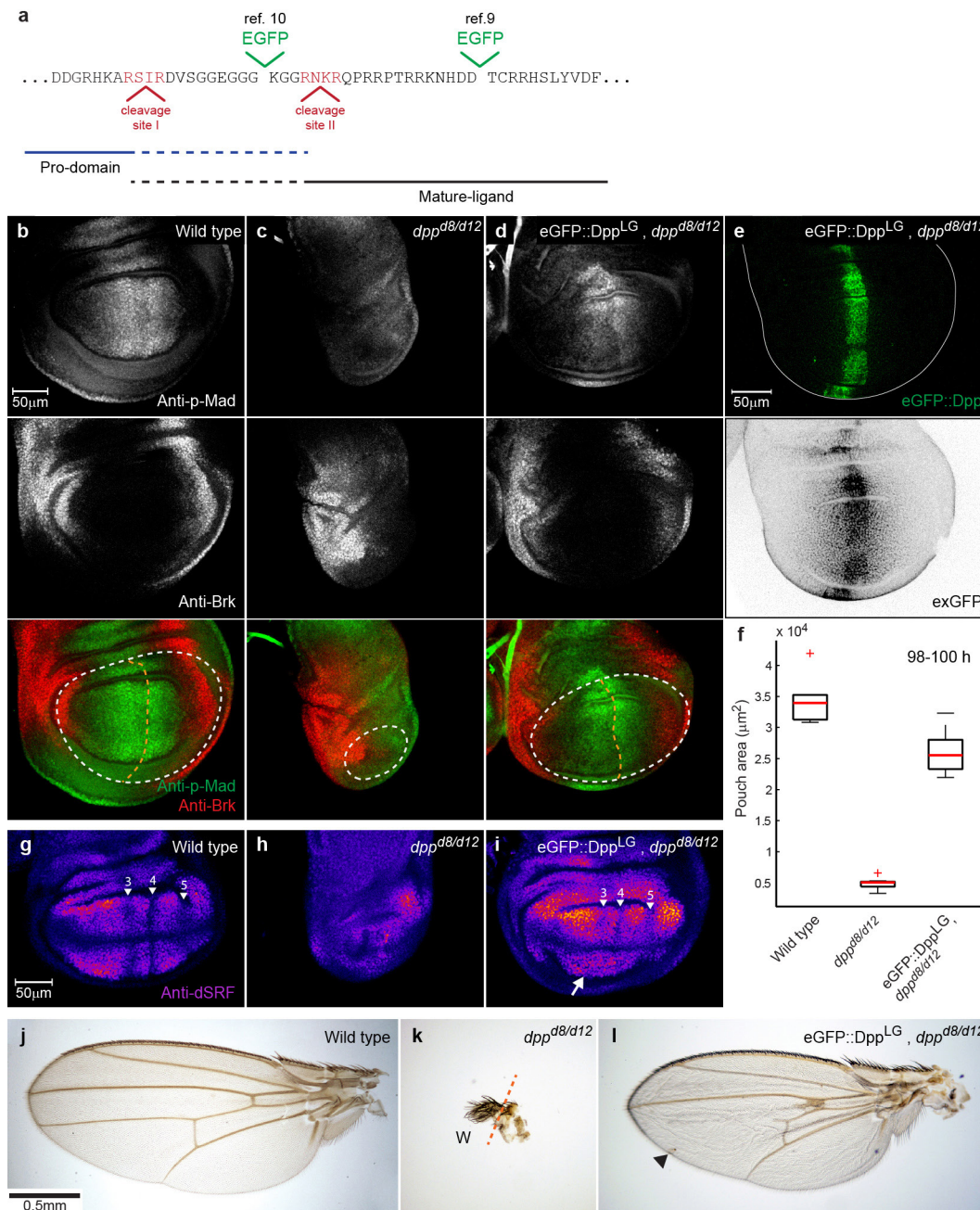
**Induction and computation of Raeppli clones.** In our experiments we used two copies of nuclear Raeppli, resulting in ten different colour combinations after induction (see ref. 43). The larvae were staged as described earlier and dissected at 96–100 h AEL. Raeppli was induced by heat shock (38°C for 15 min) at three different developmental time points: 55–59 h AEL (~41 h before dissection), 66–70 h AEL (~30 h before dissection) or 76–80 h AEL (~20 h before dissection). Discs were fixed in 4% PFA in PBS for 20 min at room temperature, washed in PBT extensively and mounted in Vectashield (H-1000, Vector Laboratories). Images were acquired on a Leica SP5 confocal microscope using the settings suggested previously<sup>43</sup>. Number of cells per clone was counted using the 'multi-point tool' in ImageJ software (National Institutes of Health). A two-sided Student's *t*-test with unequal variance was used to test for statistical significance.

**Measuring growth of the medial and lateral domain of the wing disc.** To compare the growth dynamics of the medial (high Dpp signalling) and the lateral domain (low Dpp signalling), we define the position of half-maximum Brk levels as the boundary between these two domains. The position of half-maximum Brk levels was accessed by extracting Brk intensity profiles along a straight line with 30% dorsal offset parallel to the D/V boundary (Extended Data Fig. 8f) in each disc individually. Subsequently single Brk profiles—separately for the anterior and the posterior compartment—were fit to a Hill function (see Extended Data Fig. 8f, graph 3) using the fitting-toolbox in Matlab. For fitting we excluded the lateral-most signal, which is noisy due to folds and signal from the peripodial membrane. The Hill function to which we fit the Brk profiles returns four parameters: the amplitude *A*, a measure for how sharp the profile drops *n*, a constant offset *C*, and the position of half-maximum Brk levels *k* (*k<sub>A</sub>* and *k<sub>P</sub>* for the anterior and the

posterior compartment, respectively). To access the width of the lateral domain, we measured the width of the full compartment  $L_A$  and  $L_P$  for the anterior and the posterior compartment, respectively. Since  $k_A$  equals the width of the anterior medial domain,  $L_A - k_A$  equals the width of the anterior lateral region, and accordingly  $L_P - k_P$  equals the width of the posterior lateral domain. Medial domain width in case of the posterior compartment in eGFP::Dpp morphotrap co-expressing wing discs was not fit to a Hill function, since in this condition only one cell row experiences Dpp signalling during the observed time window (equalling a width of 3.5  $\mu\text{m}$  on average).

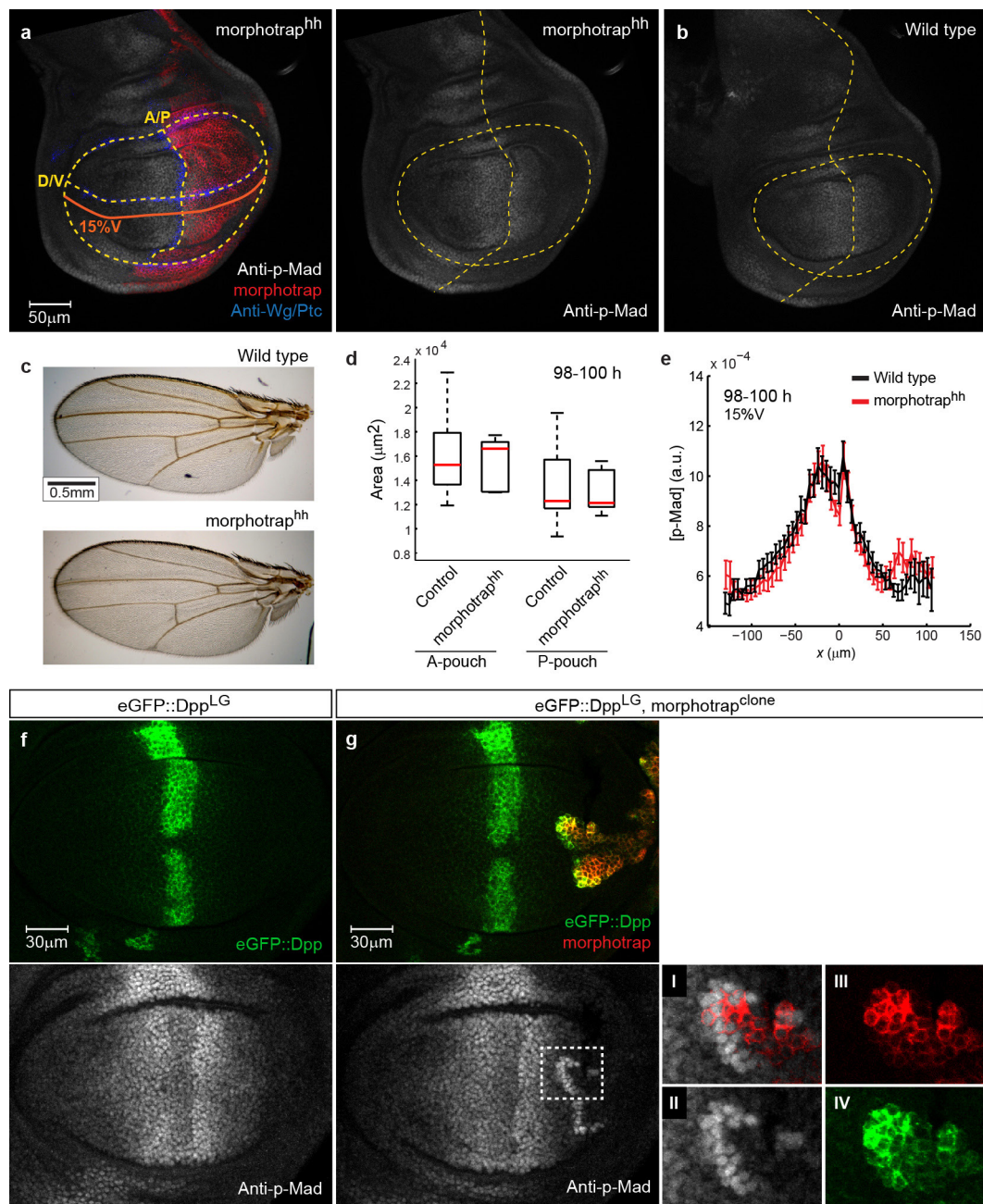
48. Yagi, R., Mayer, F. & Basler, K. Refined LexA transactivators and their use in combination with the *Drosophila* Gal4 system. *Proc. Natl Acad. Sci. USA* **107**, 16166–16171 (2010).
49. Lee, T. & Luo, L. Mosaic analysis with a repressible cell marker for studies of gene function in neuronal morphogenesis. *Neuron* **22**, 451–461 (1999).
50. Brand, A. H. & Perrimon, N. Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. *Development* **118**, 401–415 (1993).
51. Strigini, M. & Cohen, S. M. Wingless gradient formation in the *Drosophila* wing. *Curr. Biol.* **10**, 293–300 (2000).
52. Tanimoto, H., Itoh, S., ten Dijke, P. & Tabata, T. Hedgehog creates a gradient of DPP activity in *Drosophila* wing imaginal discs. *Mol. Cell* **5**, 59–71 (2000).
53. Persson, U. *et al.* The L45 loop in type I receptors for TGF- $\beta$  family members is a critical determinant in specifying Smad isoform activation. *FEBS Lett.* **434**, 83–87 (1998).
54. Kühnlein, R. P. *et al.* *spalt* encodes an evolutionarily conserved zinc finger protein of novel structure which provides homeotic gene function in the head and tail region of the *Drosophila* embryo. *EMBO J.* **13**, 168–179 (1994).
55. de Celis, J. F., Barrio, R. & Kafatos, F. C. Regulation of the *spalt/spalt-related* gene complex and its function during sensory organ development in the *Drosophila* thorax. *Development* **126**, 2653–2662 (1999).
56. Shen, J., Dahmann, C. & Pflugfelder, G. O. Spatial discontinuity of optomotor-blind expression in the *Drosophila* wing imaginal disc disrupts epithelial architecture and promotes cell sorting. *BMC Dev. Biol.* **10**, 23 (2010).
57. Schaffter, T. *From Genes to Organisms: Bioinformatics System Models and Software* (École Polytechnique Fédérale de Lausanne, 2014).
58. Künnapuu, J., Björkgren, I. & Shimmi, O. The *Drosophila* DPP signal is produced by cleavage of its proprotein at evolutionary diversified furin-recognition sites. *Proc. Natl Acad. Sci. USA* **106**, 8501–8506 (2009).
59. Foronda, D., Pérez-Garijo, A. & Martín, F. A. Dpp of posterior origin patterns the proximal region of the wing. *Mech. Dev.* **126**, 99–106 (2009).





**Extended Data Figure 1 | eGFP::Dpp can compensate for endogenous Dpp during wing disc development.** **a**, Part of the protein sequence of the Dpp protein. The two different eGFP insertion sites<sup>9,10</sup>, and the two furin cleavage sites<sup>58</sup> located in this region are marked. Furin cleavage of the inactive pro-form yields the active carboxy-terminal mature ligand. However, potential processing at cleavage site II may result in uncoupling of the eGFP from the mature ligand in the construct described previously<sup>10</sup>. We therefore inserted the EGFP C-terminal to the second furin cleavage site as was done previously<sup>9</sup>. **b–d**, Immunostainings for p-Mad and Brk in wild-type (**b**),  $dpp^{d8/d12}$  mutant (**c**) and  $dpp^{d8/d12}$  mutant wing discs rescued with eGFP::Dpp expressed under control of the  $dpp::LG^{48}$  line (**d**). In the  $dpp^{d8/d12}$  mutant wing discs expressing eGFP::Dpp, the p-Mad and Brk profiles are rescued to a control-like pattern (**d**, bottom). The pouch outline and the A/P boundary (assessed by Wg/Ptc pattern, data not shown) are marked by dotted lines. **e**, The eGFP::Dpp gradient visualized by eGFP fluorescence or by an immunostaining for the extracellular fraction of eGFP (bottom). **f**, Quantification of wing pouch area assessed by the inner Wg ring of

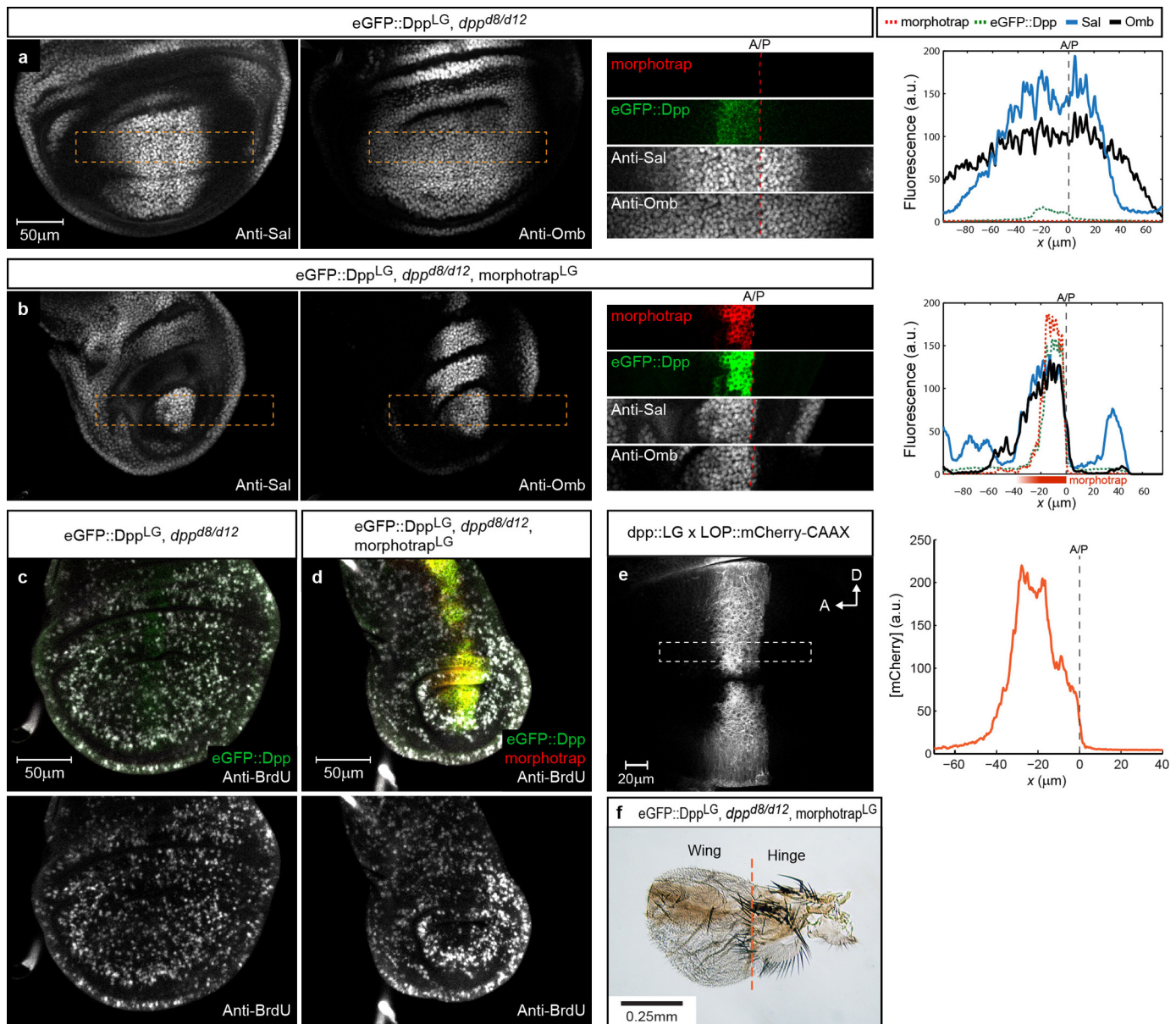
98–100 h old wing discs (wild type  $n = 6$ ,  $dpp^{d8/d12}$  mutant  $n = 10$ , rescue  $n = 10$ ; red crosses are outliers). eGFP::Dpp expression in  $dpp^{d8/d12}$  mutants rescues pouch area close to wild-type size. **g–i**, Wing discs of 98–100 h old larvae stained for *Drosophila* Serum response factor (DSRF; also known as blistered). DSRF is expressed in the future intervein tissue of the wing disc. Positions of prospective wing veins 3, 4 and 5 are marked by arrowheads. The vein pattern is largely restored in mutant discs rescued by eGFP::Dpp expression (**i**). **j–l**, Adult wings of a wild-type fly (**j**), a  $dpp^{d8/d12}$  mutant (**k**) and a  $dpp^{d8/d12}$  mutant expressing eGFP::Dpp (**l**) (W, wing). Rescued wings have a slightly elongated shape but their sizes are comparable to that of control wings. However, they show some additional vein tissue at the anterior cross-vein and wing vein 4 is absent in the distal part of the wing (marked by arrowhead). We speculate that this is due to lower eGFP::Dpp expression in the ventral compartment, which also manifests itself in lower ventral p-Mad levels (see **d**) and less well defined ventral vein patterns in the dSRF staining (**i**, arrow). Apart from these drawbacks, LexA-driven eGFP::Dpp can compensate for endogenous Dpp during wing disc development.



**Extended Data Figure 2 | Morphotrap expression does not affect growth or patterning of the wing disc.** **a**, Wing disc expressing morphotrap in the posterior compartment controlled by *hh-Gal4* (morphotrap<sup>hh</sup>). The Wg/Ptc pattern is used as a coordinate system to assess pouch size (anterior (A) pouch, left two quadrants; posterior (P) pouch, right two quadrants). Gradient profiles are measured parallel to the dorsoventral (D/V) boundary (for example, 15% ventral offset). **b**, Wild-type wing disc stained for p-Mad. **c**, Wings of a male wild-type fly and a fly expressing morphotrap in the posterior compartment under the control of *hedgehog::Gal4* (morphotrap<sup>hh</sup>). **d**, Morphotrap<sup>hh</sup> wing discs show no

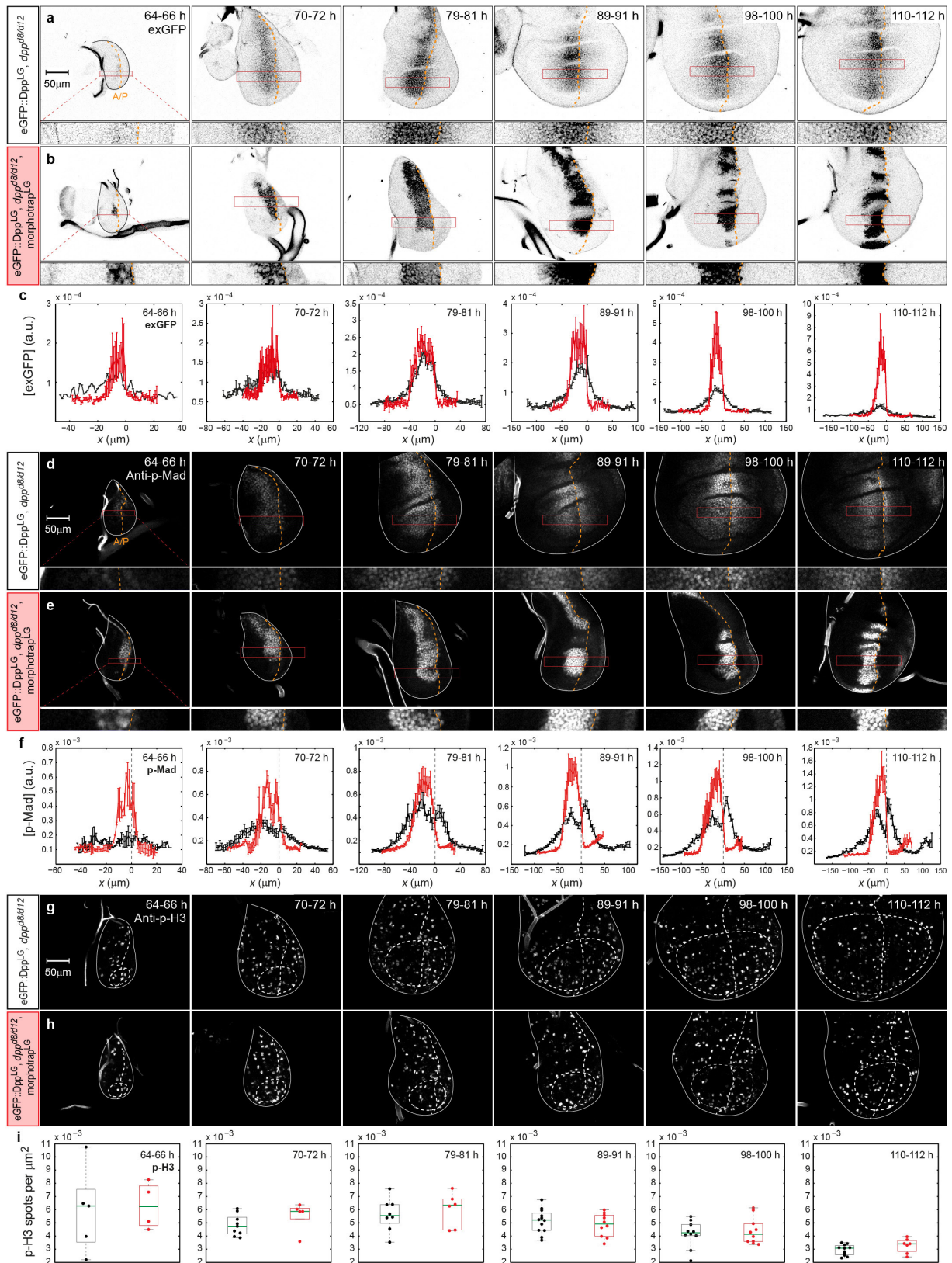
significant change in anterior or posterior pouch size (*t*-test two-sided, unequal variance: anterior compartment  $P > 0.05$ , posterior compartment  $P > 0.05$ ). **e**, Posterior expression of morphotrap does not cause obvious changes to the p-Mad profile. **f**, p-Mad pattern of a wild-type wing disc expressing eGFP::Dpp in the endogenous Dpp source area. **g**, Lateral morphotrap clones show elevated p-Mad signal at the clone boundary facing the Dpp source due to eGFP::Dpp accumulation. The region marked by a white rectangle is enlarged to the right. **d**, **e**, Control  $n = 11$ , morphotrap<sup>hh</sup>  $n = 9$ , error bars in **e** show s.d.





**Extended Data Figure 3 | Domain width of Dpp targets depends on Dpp spreading.** **a**, Discs of a *dpp<sup>d8/d12</sup>* mutant rescued with *eGFP::Dpp* stained for Dpp targets Sal and Omb. Omb shows a wider distribution than Sal. **b**, *dpp<sup>d8/d12</sup>* mutant wing discs co-expressing *eGFP::Dpp* and morphotrap. The regions marked by a dotted rectangle are enlarged to the right of the respective image. The dotted red line marks the A/P compartment boundary. In the absence of Dpp spreading, target domains collapse onto a single cell row in the posterior compartment. In the anterior compartment domain borders are less sharp. We hypothesize that this is due to morphotrap-bound *eGFP::Dpp* that is dragged into the anterior compartment by dividing cells (see also **e**). Intensity profiles of the enlarged regions are plotted to the right. **c**, Wing disc of a *dpp<sup>d8/d12</sup>* mutant rescued with *eGFP::Dpp* stained for the proliferation marked BrdU. Uniform BrdU signal is obtained along the entire disc tissue.

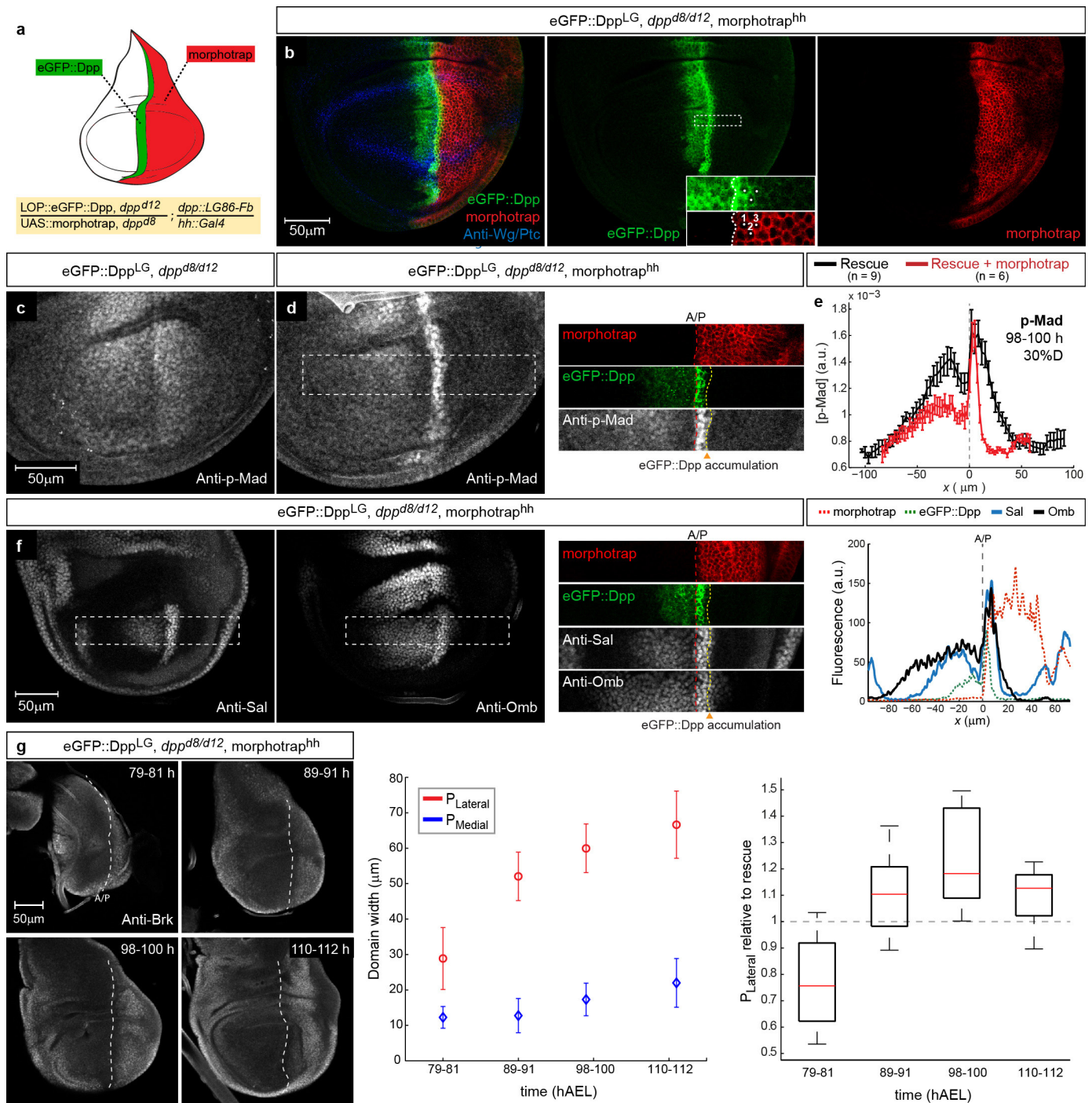
**d**, Rescued wing disc with blocked Dpp spreading stained for BrdU. Also in the absence of Dpp spreading the uniform BrdU signal is not lost. **e**, Expression of mCherry-CAAX under the control of the *dpp::LexA* driver line used for the rescue. mCherry-CAAX is a protein with a long half-life that localizes to the membrane. The graph to the right shows intensity plot of the region marked on the left. No posterior expression is observed; however, the protein profile is graded into the anterior compartment. Analogous to morphotrap-bound *eGFP::Dpp*, the stable mCherry-CAAX protein forms a concentration gradient into the anterior compartment due to dividing cells that are pushed further laterally into the anterior compartment. **f**, Wing of a rescued fly with blocked Dpp spreading. The hinge region, arising from the lateral wing disc region, is present and well patterned. In contrast, the wing field, arising from the medial wing disc region, is strongly reduced in size and patterning is lost.



**Extended Data Figure 4 | Time course of eGFP::Dpp spreading, signalling and the mitotic index. a–i,** Time course of extracellular eGFP::Dpp (exGFP), Dpp signalling (p-Mad) and p-H3 from 64–112 h AEL of larval development. **a, b,** Representative discs of the six time points examined of control animals (**a**) and animals with blocked Dpp spreading (**b**) stained for exGFP. The region marked by a red rectangle is enlarged below each image. eGFP::Dpp spreading is tightly blocked by morphotrap at all time points. **c,** Average exGFP profiles for all time points (control in black/block in red:  $n = 43/29$ ). **d, e,** Discs of control animals (**d**)

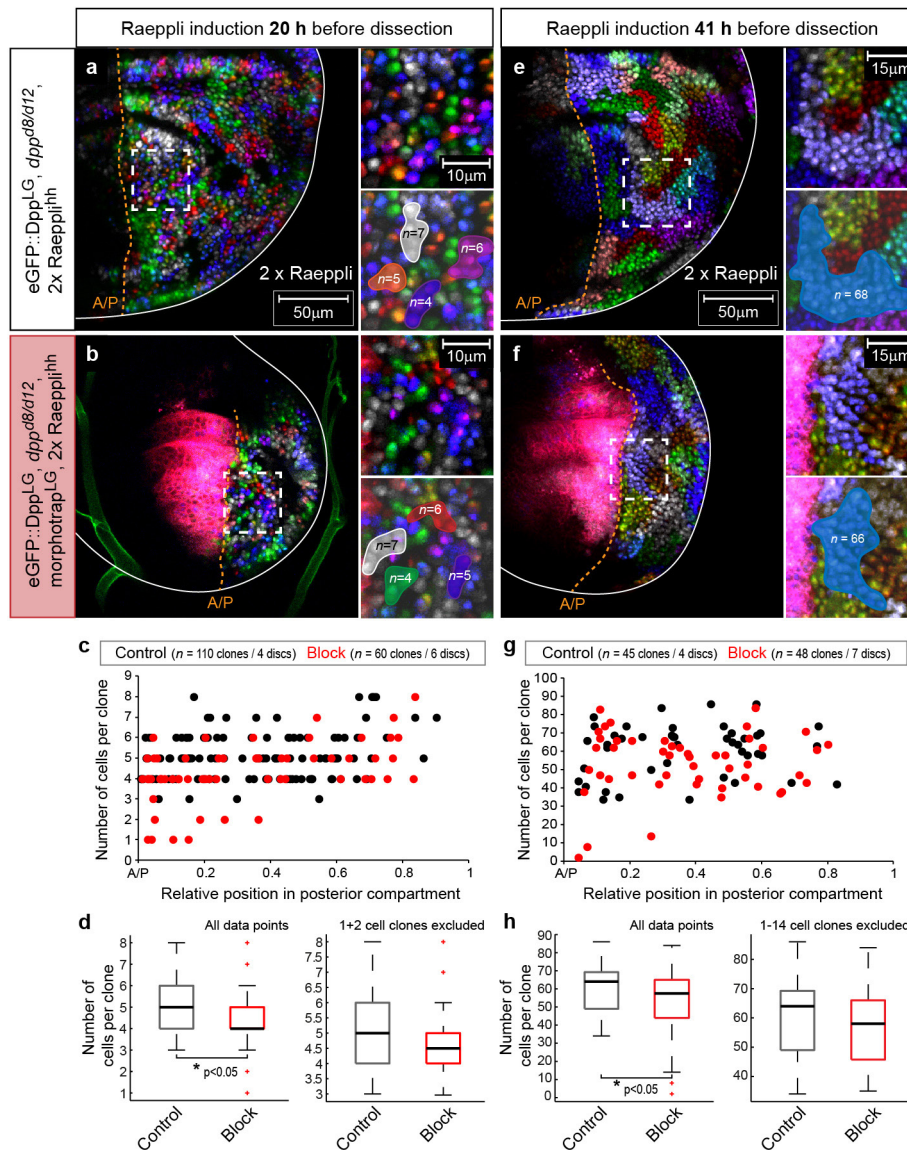
and animals with blocked Dpp spreading (**e**) stained for p-Mad. When Dpp spreading is blocked, the p-Mad gradient also collapses onto the source region at all time points. **f,** Average p-Mad profiles (control/block:  $n = 50/35$ ). **g, h,** Control discs (**g**) and discs with blocked Dpp spreading (**h**) stained for p-H3. **i,** Quantification of the mitotic index (p-H3 spot density). No significant differences were observed between control discs (black,  $n = 55$ ) and discs with blocked Dpp spreading (red,  $n = 43$ ) at any time point ( $n > 0.05$  for all time points, two-sided  $t$ -test, unequal variance).





**Extended Data Figure 5 | Shortening of the Dpp gradient by posterior morphotrap expression.** **a**, Scheme of morphotrap expression in the posterior compartment (using *hh::Gal4*) in *dpp<sup>d8/d12</sup>* mutant wing discs rescued with eGFP::Dpp. **b**, Posterior morphotrap expression in the rescue background results in strong eGFP signal in the first three cell rows of the posterior compartment due to eGFP::Dpp accumulation; after three cell rows the eGFP fluorescence signal drops. **c**, p-Mad staining in a *dpp<sup>d8/d12</sup>* mutant wing disc rescued with eGFP::Dpp. **d**, p-Mad staining in a *dpp<sup>d8/d12</sup>* mutant wing disc rescued by eGFP::Dpp and expressing morphotrap in the posterior compartment. Note that the eGFP::Dpp accumulation (marked by a yellow line) directly overlaps with the observed p-Mad signal. **e**, The average p-Mad profiles show that the p-Mad gradient range directly depends on the range of Dpp spreading (error bars are s.d.). **f**, *dpp<sup>d8/d12</sup>* mutant wing discs rescued with eGFP::Dpp expressing morphotrap in the posterior compartment stained for Sal and Omb (for control discs

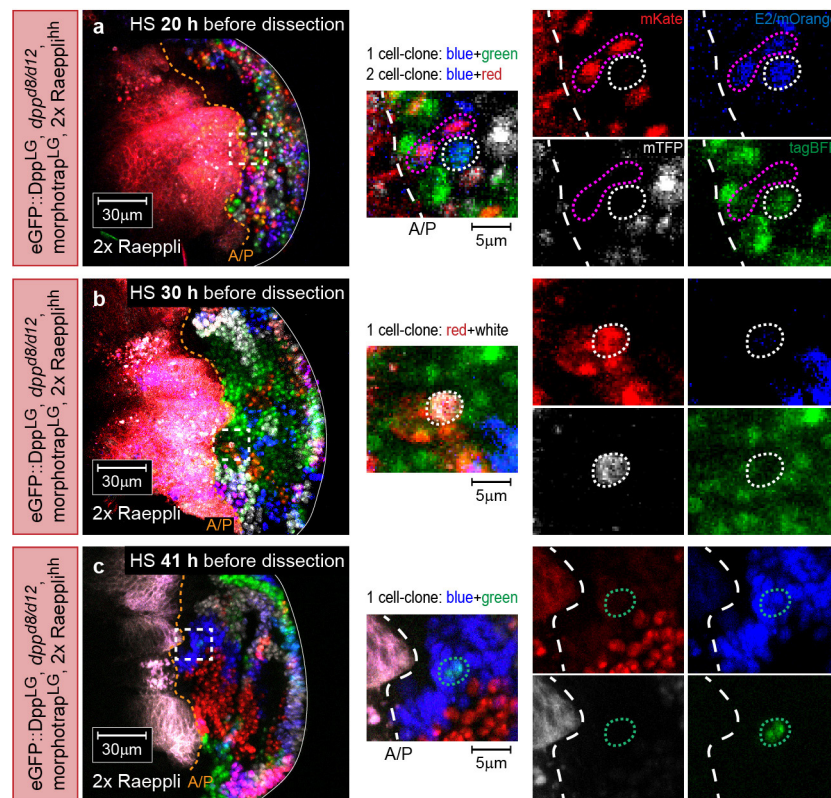
see Extended Data Fig. 3a). The A/P boundary is marked by a dotted red line and the range of the eGFP::Dpp accumulation is marked by a dotted yellow line. In this condition the domain width of both targets is strongly reduced. The Sal domain directly collapses onto the eGFP::Dpp accumulation domain. However, Omb, which can be activated at lower Dpp signalling levels, shows a slightly wider distribution. We hypothesize that this is again due to morphotrap-stabilized eGFP::Dpp being dragged into the posterior compartment (as discussed in Extended Data Fig. 3). Intensity profiles of the enlarged regions are plotted to the right. **g**, Representative *dpp<sup>d8/d12</sup>* mutant wing discs rescued with eGFP::Dpp expressing morphotrap in the posterior compartment stained for Brk at the indicated time points (79-12 h AEL). In this condition the medial region shows strongly reduced growth (compare to Fig. 5a-f). However, the growth dynamics of the lateral domain are similar to the lateral growth observed in control wing discs (right).



**Extended Data Figure 6 | Clonal growth rates do not change in the absence of Dpp spreading.** **a–h**, Estimation of clonal proliferation rates as shown in Fig. 4, inducing *Raepl<sup>ih</sup>* at different time points: either 20 h before dissection (**a–d**) or 41 h before dissection (**e–h**). Discs were dissected at 96–100 h AEL. **a, e**, Representative control discs. **b, f**, Representative discs with blocked Dpp spreading. **c, g**, Clone size (number of cells per clone) plotted against the relative position in the posterior compartment (0 corresponding to the A/P boundary and 1 to

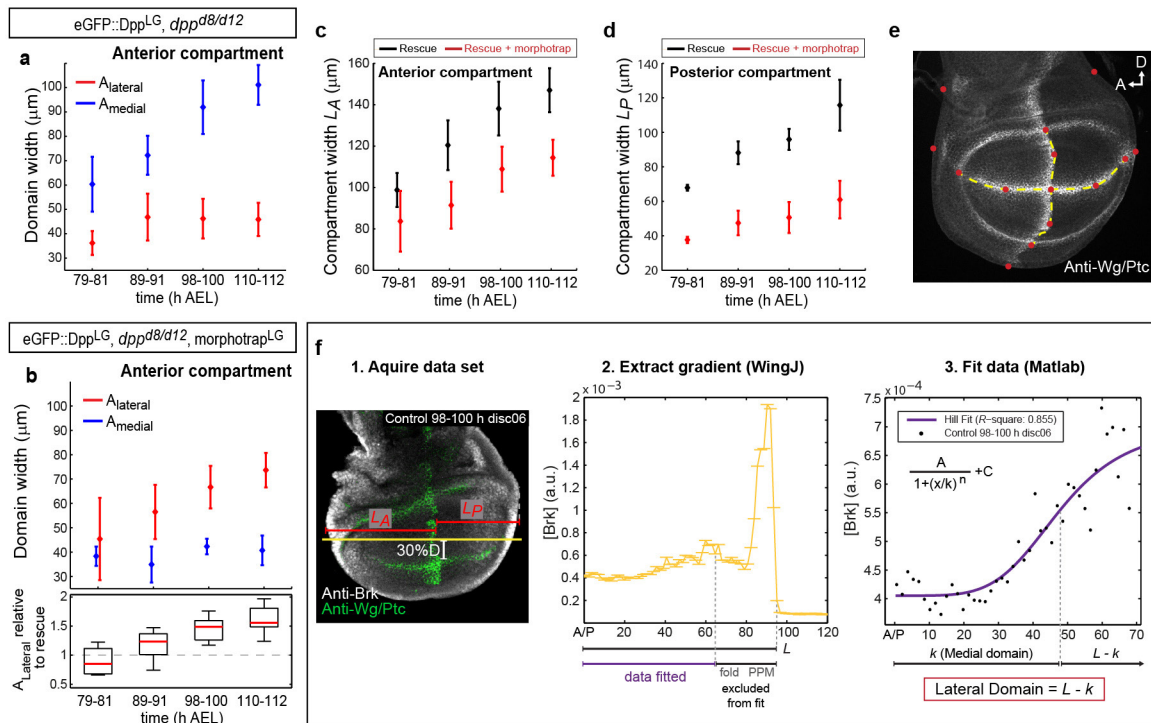
the posterior edge of the disc). Low numbers of small clones in proximity to the A/P boundary are found in discs with blocked Dpp spreading (red dots), while these small clones are not present in control discs (black dots; see also Extended Data Fig. 7). **d, h**, Boxplots showing the number of cells per clone. When the small clones are excluded (right boxplots) no significant differences are detected in clonal proliferation between control discs and discs with blocked Dpp spreading ( $P > 0.05$ ).





**Extended Data Figure 7 | Small clones in discs with blocked Dpp spreading.** **a–c**, Wing discs with blocked Dpp spreading carrying small Raeppli clones in proximity of the A/P boundary. Raeppli was induced by

heat shock (HS) at different time points during larval development: 20 h (**a**), 30 h (**b**) and 41 h (**c**) before dissection. The regions marked by a white rectangle in the left column are magnified to the right.

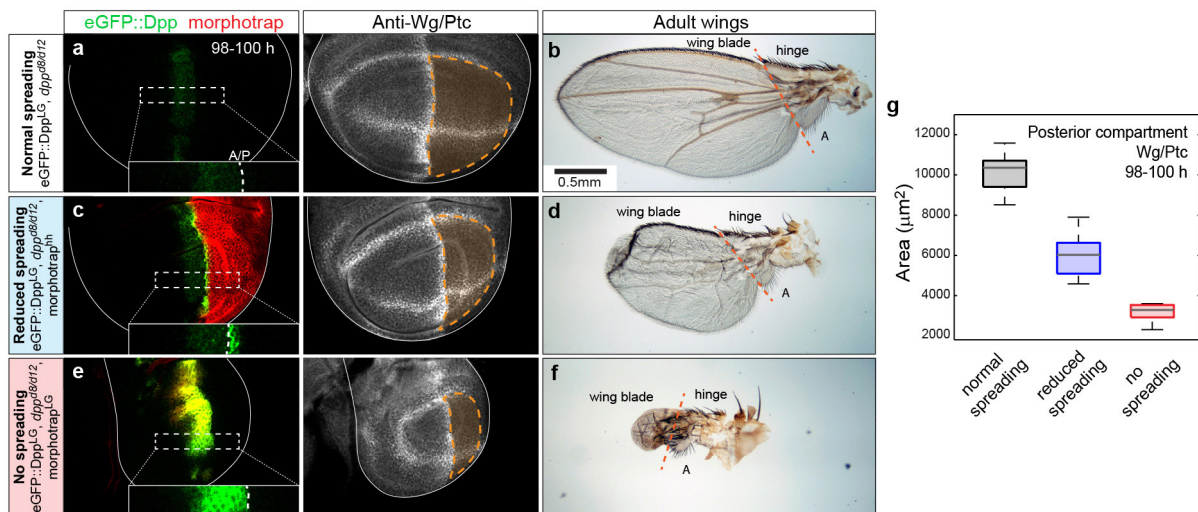


#### Extended Data Figure 8 | Temporal development and fitting procedure of Brk data set.

**a, b**, As can be seen in Fig. 2f, there is a gap in Brk expression in the lateral-most region of the posterior compartment, indicative of Dpp expression from another, laterally located source. Indeed, it has been shown that Dpp is expressed during the third instar larval stage in a posterior, lateral position and exerts a patterning role on the wing imaginal disc. However, this late Dpp expression does not affect the growth properties of wing disc cells<sup>59</sup>. Despite this, the additional Dpp source might complicate the interpretation of our growth analyses. To circumvent this problem, we also measured the growth properties in the anterior compartment in the presence (**a**) and in the absence of the eGFP::Dpp gradient (**b**; high uniform levels of Brk are indeed present in all cells outside the source). Indeed, we found that the lateral anterior region still grows despite the absence of the Dpp gradient and the lack of Dpp signalling. **c, d**, Width of the anterior and posterior compartment

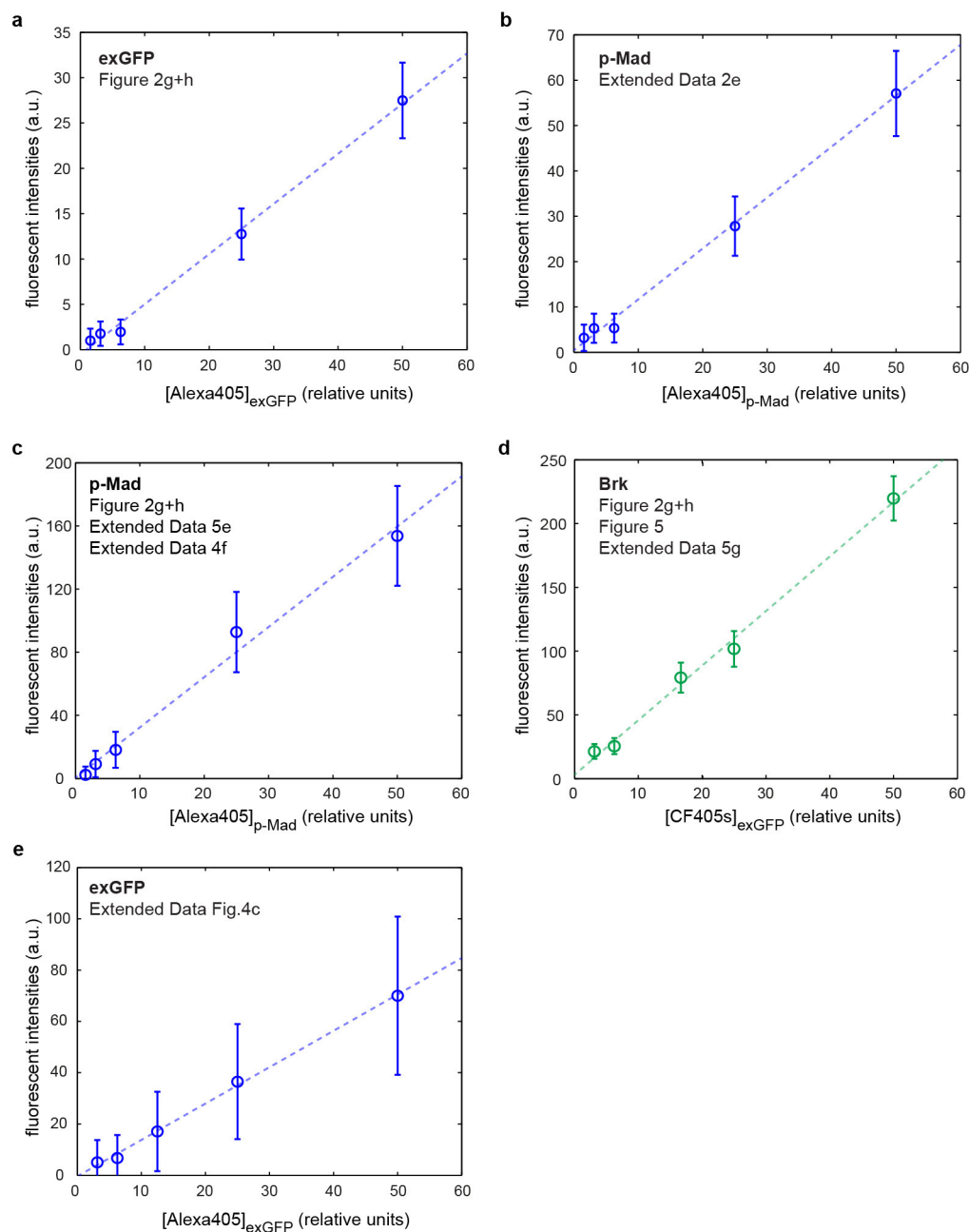
respectively in *dpp*<sup>d8/d12</sup> mutant wing discs rescued with eGFP::Dpp (black,  $n = 34$ ) and *dpp*<sup>d8/d12</sup> mutant wing discs co-expressing eGFP::Dpp and morphotrap (red,  $n = 37$ , error bars show s.d.). **e**, The red dots mark the 15 points used as landmarks for the affine transformation. Using affine transformation allowed us to overlay discs of slightly different shapes and sizes when generating the mitotic density maps shown in Fig. 3c, e (see also Methods for details). **f**, Computation of Brk data set shown for the posterior compartment: (1) The compartment width  $L_A$  or  $L_P$  was defined as the distance from the A/P boundary to the anterior or posterior edge of the wing tissue, respectively. Brk profiles were measured along a straight line with 30%D offset. (2) Profiles were extracted using WingJ software. (3) The single gradients were fitted to the shown Hill function. The fitting procedure returns the parameter  $k$ , which corresponds to the position of half-maximum Brk levels and hence to the width of the medial domain. Therefore, the lateral domain equals  $L - k$ .





**Extended Data Figure 9 | Impact of Dpp spreading on wing pouch and adult wing size.** **a**, Dpp mutant wing disc rescued with eGFP::Dpp stained for Wg (outlining the wing pouch) and Ptc (marking the A/P boundary). In this background eGFP::Dpp spreading is not hindered and a normal gradient forms. The size of the posterior wing pouch is estimated by the area enclosed by the Wg ring and the A/P boundary (coloured orange) and plotted in **g**. **b**, Adult wing of a rescued fly. The border between the hinge region and the wing blade is marked by a dotted orange line; the alula is labelled with an A. (Wing is the same as shown in Extended Data Fig. 11.) **c**, Rescued wing disc expressing morphotrap in the posterior compartment, reducing Dpp dispersal range in the posterior compartment. In this condition pouch size is significantly decreased (see **g**). **d**, Wing of a rescued fly expressing morphotrap in the posterior

compartment. The wing blade area is strongly decreased and patterning in the posterior part of the wing is lost. **e**, Rescued wing disc expressing morphotrap in the Dpp stripe, completely blocking Dpp spreading, and hence gradient formation. Full block of Dpp spreading results in a further decrease of the Wg/Ptc-encircled posterior pouch area. **f**, Wing of a rescued fly co-expressing eGFP::Dpp and morphotrap. Full block of Dpp spreading results in a strong reduction of wing blade area. Only a small amount of unpatterned wing tissue is left, while the hinge region seems to be patterned normally (alula is present). **g**, Plot of the posterior pouch area, as accessed by the Wg/Ptc staining shown in (**a**, **c**, **e**, right) when Dpp spreads normally (black), Dpp spreading is reduced (blue) or when Dpp spreading is fully blocked (red). With decreasing Dpp dispersal range also the posterior pouch area decreases ( $n = 22$ ).



**Extended Data Figure 10 | Linear range imaging conditions.** a–e, Linear range imaging for the quantitative data sets acquired (corresponding figure is labelled at top left in each plot). Dilutions of the secondary antibodies used (anti-rb-Alexa 405 (blue) and anti-gp-CF405S (green)) in Vectashield mounting medium yield fluorescent intensities proportional

to their concentrations under the established imaging conditions. Mean intensities were extracted using the Histogram function in ImageJ on the whole imaging field of a mean projection. The background fluorescence was measured by imaging a slide only containing Vectashield and subtracted from the mean values. Dotted lines indicate linear fits.



# Novel antibody–antibiotic conjugate eliminates intracellular *S. aureus*

Sophie M. Lehar<sup>1</sup>, Thomas Pillow<sup>2</sup>, Min Xu<sup>3</sup>, Leanna Staben<sup>2</sup>, Kimberly K. Kajihara<sup>1</sup>, Richard Vandlen<sup>4</sup>, Laura DePalatis<sup>4</sup>, Helga Raab<sup>4</sup>, Wouter L. Hazenbos<sup>1</sup>, J. Hiroshi Morisaki<sup>1</sup>, Janice Kim<sup>3</sup>, Summer Park<sup>3</sup>, Martine Darwish<sup>4</sup>, Byoung-Chul Lee<sup>4</sup>, Hilda Hernandez<sup>5</sup>, Kelly M. Loyet<sup>5</sup>, Patrick Lupardus<sup>6</sup>, Rina Fong<sup>6</sup>, Donghong Yan<sup>3</sup>, Cecile Chalouni<sup>7</sup>, Elizabeth Luis<sup>4</sup>, Yana Khalfin<sup>5</sup>, Emile Plise<sup>8</sup>, Jonathan Cheong<sup>8</sup>, Joseph P. Lyssikatos<sup>2</sup>, Magnus Strandh<sup>9</sup>, Klaus Koefoed<sup>9</sup>, Peter S. Andersen<sup>9</sup>, John A. Flygare<sup>2</sup>, Man Wah Tan<sup>1</sup>, Eric J. Brown<sup>1</sup> & Sanjeev Mariathasan<sup>1</sup>

***Staphylococcus aureus* is considered to be an extracellular pathogen. However, survival of *S. aureus* within host cells may provide a reservoir relatively protected from antibiotics, thus enabling long-term colonization of the host and explaining clinical failures and relapses after antibiotic therapy. Here we confirm that intracellular reservoirs of *S. aureus* in mice comprise a virulent subset of bacteria that can establish infection even in the presence of vancomycin, and we introduce a novel therapeutic that effectively kills intracellular *S. aureus*. This antibody–antibiotic conjugate consists of an anti-*S. aureus* antibody conjugated to a highly efficacious antibiotic that is activated only after it is released in the proteolytic environment of the phagolysosome. The antibody–antibiotic conjugate is superior to vancomycin for treatment of bacteraemia and provides direct evidence that intracellular *S. aureus* represents an important component of invasive infections.**

*S. aureus* is the leading cause of bacterial infections in humans worldwide and represents a major health problem in both hospital and community settings<sup>1</sup>. However, *S. aureus* is not exclusively a pathogen and commonly colonizes the anterior nares and skin of healthy individuals. When infection does occur, the most serious infections such as endocarditis, osteomyelitis, necrotizing pneumonia and sepsis take hold after dissemination of the bacteria into the bloodstream<sup>2</sup>. Over the last several decades, infection with *S. aureus* has become increasingly difficult to treat due to the emergence and rapid spread of methicillin-resistant *S. aureus* (MRSA), which is resistant to all known  $\beta$ -lactam antibiotics<sup>3</sup>. Alarming, reduced susceptibility to vancomycin and resistance to linezolid and daptomycin have already been reported in MRSA clinical strains<sup>4</sup>.

Investigations going back at least 50 years have revealed that *S. aureus* is able to invade and survive inside mammalian cells, including the phagocytic cells that are responsible for bacterial clearance<sup>5–11</sup>. *S. aureus* is taken up by host phagocytic cells, primarily neutrophils and macrophages, within minutes after intravenous infection<sup>12</sup>. While the majority of the bacteria are effectively killed by these cells, incomplete clearance of *S. aureus* inside blood-borne phagocytes can allow these infected cells to act as ‘Trojan horses’ for dissemination of the bacteria away from the initial site of infection. Indeed, patients with normal neutrophil counts may be more prone to disseminated disease than those with reduced neutrophil counts<sup>5,13,14</sup>. Once delivered to the tissues, *S. aureus* can invade various non-phagocytic cell types, and intracellular *S. aureus* in tissues is associated with chronic or recurrent infections including osteomyelitis<sup>15</sup>, recurrent rhinosinusitis<sup>16</sup>, pulmonary infections<sup>17</sup> and endocarditis<sup>18</sup>. In addition, intracellular *S. aureus* has been shown to undermine innate immune responses and induce subsequent destruction of neutrophils<sup>19,20</sup>. Together, these data suggest that ablating intracellular *S. aureus* is key to clinical success, but until now there has been no way to test this hypothesis directly.

## Intracellular MRSA are protected from antibiotics

To confirm the hypothesis that mammalian cells provide a protective niche for *S. aureus* in the presence of antibiotic therapy, we compared the efficacy of three major antibiotics that are currently used as the standard of care for invasive MRSA infections (vancomycin, daptomycin and linezolid) against extracellular planktonic bacteria versus bacteria sequestered inside murine macrophages (Table 1). All three antibiotics failed to kill a highly virulent community-acquired MRSA strain, USA300, sequestered inside macrophages exposed to clinically achievable concentrations of the antibiotics, consistent with other studies showing that the majority of existing antibiotics are inefficient at killing intracellular *S. aureus* both *in vitro*<sup>21</sup> and *in vivo*<sup>22</sup>.

## Intracellular *S. aureus* spreads infection

To compare directly the virulence of intracellular bacteria versus free-living planktonic bacteria and to determine whether the intracellular bacteria are able to establish infection in the presence of vancomycin *in vivo*, mice were infected with equivalent doses of *S. aureus* taken directly from broth culture or bacteria sequestered inside host peritoneal

**Table 1 | Antibiotic minimum inhibitory concentrations for MRSA**

Antibiotics	Extracellular MRSA (MIC ( $\mu\text{g ml}^{-1}$ ))	Intracellular MRSA (MIC ( $\mu\text{g ml}^{-1}$ ))	Serum $C_{\text{max}}$ ( $\mu\text{g ml}^{-1}$ )
Vancomycin	1	>100	50
Daptomycin	4	>100	60
Linezolid	0.3	>20	20
Rifampicin	0.004	50	20

Extracellular MIC is the minimum antibiotic dose that prevented growth of MRSA seeded in trypticase soy broth. Intracellular MIC is the minimum dose that prevented survival of MRSA sequestered inside murine macrophages (> indicates continued growth at the highest antibiotic dose tested). Data are summarized from three different experiments. Serum  $C_{\text{max}}$ , the expected serum concentrations for clinically relevant antibiotics<sup>37</sup>.

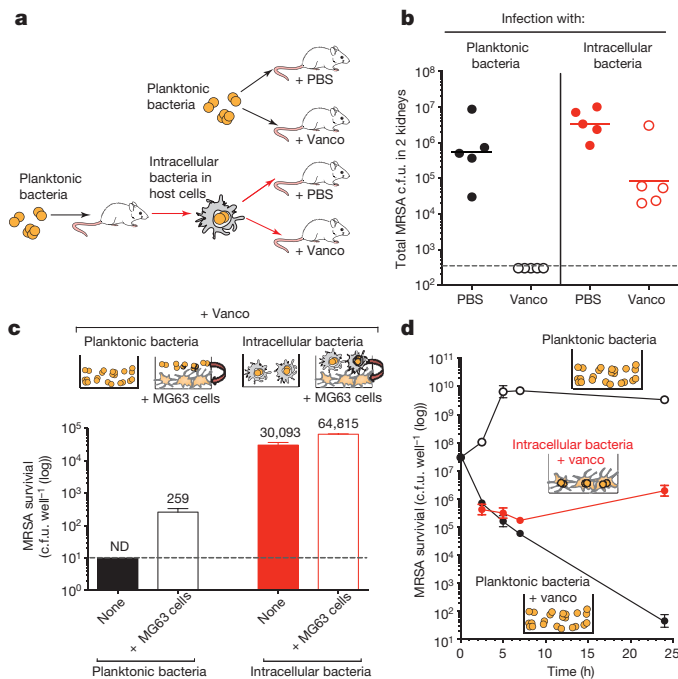
<sup>1</sup>Infectious Diseases Department, Genentech Inc., South San Francisco, California 94080, USA. <sup>2</sup>Medicinal Chemistry Department, Genentech Inc., South San Francisco, California 94080, USA.

<sup>3</sup>Translational Immunology Department, Genentech Inc., South San Francisco, California 94080, USA. <sup>4</sup>Protein Chemistry Department, Genentech Inc., South San Francisco, California 94080, USA.

<sup>5</sup>Biochemical and Cellular Pharmacology Department, Genentech Inc., South San Francisco, California 94080, USA. <sup>6</sup>Structural Biology Department, Genentech Inc., South San Francisco, California 94080, USA.

<sup>7</sup>Pathology Department, Genentech Inc., South San Francisco, California 94080, USA. <sup>8</sup>Drug metabolism and Pharmacokinetics Department, Genentech Inc., South San Francisco, California 94080, USA.

<sup>9</sup>Symphogen A/S, Pederstrupvej 93, DK-2750 Ballerup, Denmark.



**Figure 1 | Intracellular MRSA are protected from vancomycin.**

**a**, Experimental design for generating planktonic versus intracellular bacteria for infection and treatment with vancomycin (vanco). **b**, Bacterial loads in kidney, 4 days after infection. c.f.u., colony-forming units. Each point represents data from a single mouse ( $n = 5$  mice per group). **c**, Planktonic or intracellular bacteria were suspended in media containing vancomycin and cultured alone or on a monolayer of MG63, and surviving bacteria were enumerated 1 day later. Error bars represent means  $\pm$  standard deviation (s.d.). **d**, Survival of MRSA cultured with vancomycin with or without MG63 cells. Error bars show s.d. ND, none detected. Dashed lines indicate the limit of detection. Representative of three independent experiments.

macrophages and neutrophils (Fig. 1a). Mice infected with intracellular bacteria had equivalent or higher bacterial burdens in the kidneys 4 days after infection compared with those infected with planktonic bacteria (Fig. 1b). Infection with intracellular bacteria also resulted in more consistent colonization of the brain (Extended Data Fig. 1). To characterize this observation further, we infected MG63 osteoblasts with either planktonic MRSA or an equivalent number of intracellular MRSA, in the presence of vancomycin. Planktonic bacteria exposed to vancomycin alone were efficiently killed, (Fig. 1c). However, we recovered a small number of surviving bacteria (approximately 0.06% of input) associated with the MG63 cells 1 day after infection, which had been protected from vancomycin by invasion into osteoblasts. MRSA that were sequestered inside peritoneal cells showed a significant increase in both survival and efficiency of MG63 infection in the presence of vancomycin. About 15% of intracellular MRSA in the leukocytes survived under vancomycin exposure identical to that which sterilized cultures of planktonic bacteria. Intracellular *S. aureus* that were sequestered in MG63 cells (Fig. 1d), primary human brain endothelial cells and A549 bronchial epithelial cells (Extended Data Fig. 2) were also able to increase by almost tenfold over a 24 h period under constant exposure to a concentration of vancomycin that killed free-living bacteria. Together, these data suggest that intracellular reservoirs of MRSA in myeloid cells can promote dissemination of infection to new sites in the presence of active antibiotic treatment, and that intracellular growth can occur in a variety of cell types despite constant antibiotic therapy.

## Designing the antibody–antibiotic conjugate

These findings suggested that therapies aimed at eliminating intracellular bacteria may improve clinical success. To test this conclusively,

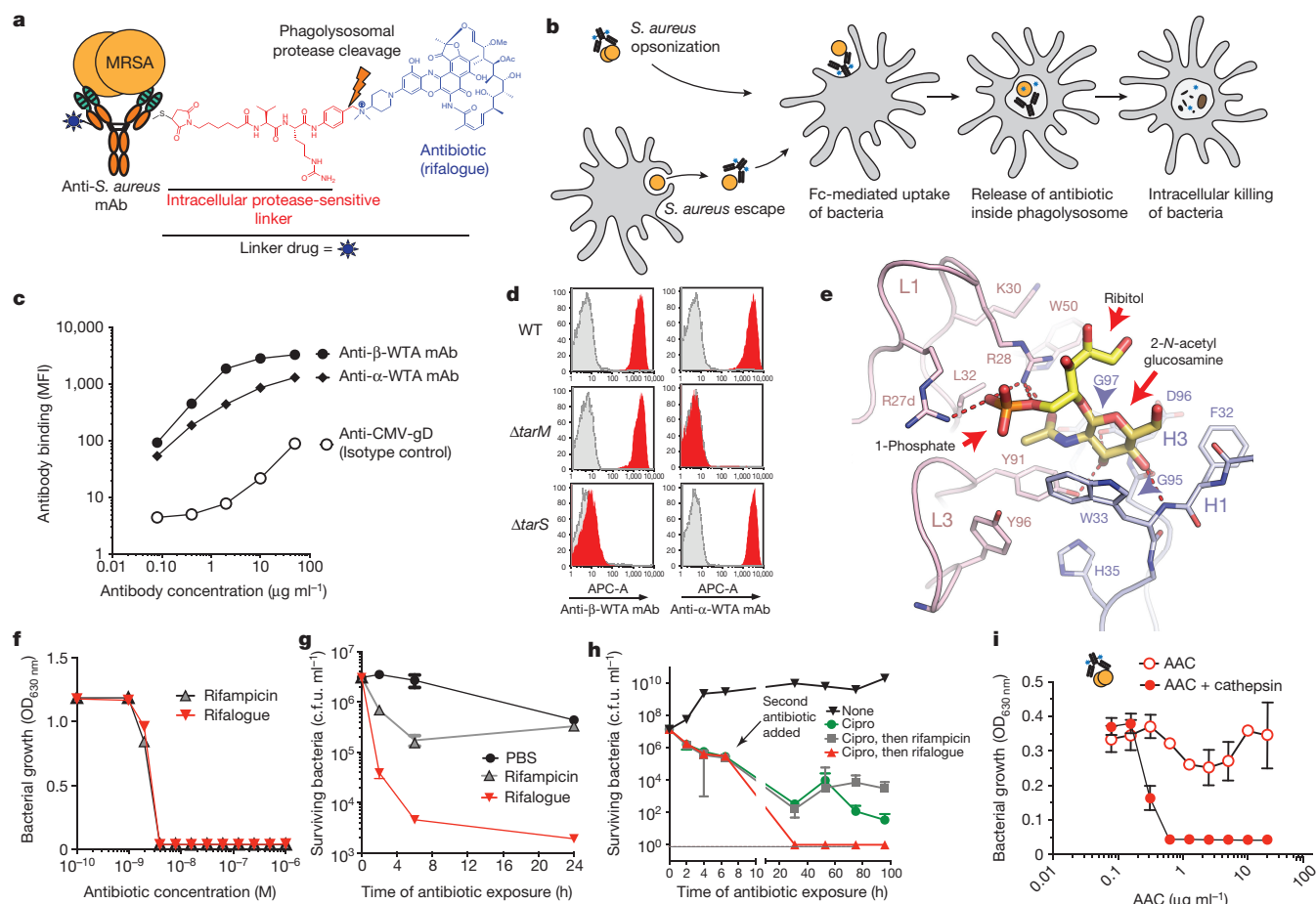
we developed a novel therapeutic that is activated specifically inside mammalian cells that have taken up *S. aureus*. The antibody–antibiotic conjugate (AAC) consists of an anti-*S. aureus* antibody (THIOMAB, engineered to contain unpaired cysteines) covalently linked via the introduced cysteines to a highly efficacious antibiotic using a cathepsin-cleavable linker containing a novel quaternary ammonium salt (Fig. 2a). The resulting AAC has no direct antibacterial activity when bound to planktonic *S. aureus* and does not diffuse into mammalian cells, owing to the size of the antibody. However, when AAC-opsonized bacteria are taken up by host cells, intracellular proteases cleave the linker and readily release the antibiotic in its active form<sup>23</sup> (Fig. 2b). Depending on the antibody specificity, we found that thousands of AACs can bind to a single bacterium and deliver a sufficient concentration of free antibiotic within the phagosome to result in bacterial killing.

The antibody and antibiotic components were carefully chosen and optimized for maximal efficacy. A panel of more than 40 anti-*S. aureus* antibodies were cloned and purified from B cells derived from the peripheral blood of patients recovering from various *S. aureus* infections and screened for binding to a panel of clinically relevant *S. aureus* strains and to USA300 isolated directly *ex vivo* from the kidneys of infected mice. We found the greatest extent of binding with antibodies directed against wall-teichoic acids (WTAs), pathogen-specific polyanionic glycopolymers that are connected to the thick peptidoglycan layers of Gram-positive bacteria (Fig. 2c). *S. aureus* produces WTAs composed of phospho-ribitol repeating units that are further modified by either  $\alpha$ - or  $\beta$ -O-linked *N*-acetylglucosamine (GlcNAc) sugars mediated by TarM or TarS glycosyltransferases, respectively<sup>24</sup>. A human immunoglobulin G<sub>1</sub> (IgG<sub>1</sub>) that recognizes  $\beta$ -O-linked GlcNAc sugar modifications on WTA bound to all *S. aureus* strains tested. Monoclonal antibodies recognizing the  $\alpha$ -O-linked GlcNAc bound well to *S. aureus* strains cultured *in vitro*; however, expression of the  $\alpha$ -O-linked GlcNAc was absent on some *S. aureus* isolates. On those strains that co-expressed the  $\alpha$ - and  $\beta$ -O-linked GlcNAc,  $\beta$ -specific antibodies yielded consistently higher binding to *in vivo*-derived MRSA (Fig. 2c). Antigen specificity of the antibodies was confirmed by genetic means, such that antibodies against  $\alpha$ - or  $\beta$ -GlcNAc sugar modifications on WTAs failed to bind to *S. aureus* strains lacking the TarM or TarS glycosyltransferase, respectively (Fig. 2d). We estimate that  $\sim 15,000$   $\beta$ -WTA-specific antibodies can bind to protein-A-deficient *S. aureus* USA300 by measuring binding of radiolabelled antibodies (data not shown).

To characterize antibody binding to WTA at the molecular level further, we co-crystallized the Fab fragment from an anti- $\beta$ -WTA antibody in complex with a synthetic form of the minimal repeating  $\beta$ -WTA unit (C2 carbon of ribitol-1-phosphate linked to the C1 position in GlcNAc via a  $\beta$ -glycosidic bond) and determined a crystal structure of the complex at 1.7 Å resolution (Extended Data Table 1). The structure reveals how the minimal  $\beta$ -WTA epitope binds to the complementarity determining region (CDR) of the Fab antibody fragment (Fig. 2e). The heavy chain H1 and H3 CDR loops form a binding site for the GlcNAc, with the GlcNAc hexose ring stacking against the Trp 33 indole side chain. Importantly, the antibody has an extended L1 loop containing two arginine residues (Arg 27d and 28) that interact with the phosphate group on the ribitol backbone. This arginine ‘tweezers’ motif triangulates the ribitol phosphodiester backbone in relation to the GlcNAc moiety, and is probably a key reason for the  $\beta$ -anomer-specific recognition of GlcNAc-WTA by the antibody.

An extensive set of criteria was used to choose the optimal antibiotic for the AAC platform. We chose the rifamycin class of antibiotics for their high potency, unaltered bactericidal activity in low phagolysosomal pH, and their ability to withstand intracellular insults. A rifamycin derivative (a rifalogue) that allowed connection to the linker antibody through a tertiary amine (T. Pillow, manuscript in preparation) had optimum physicochemical properties to allow for efficient bacterial killing. The chemical groups known to make direct contact with





**Figure 2 | AAC design.** **a**, Model of AAC (not drawn to scale). **b**, Mechanism of AAC action. **c**, Binding of Alexa-488 anti- $\beta$ -GlcNAc WTA monoclonal antibody (mAb) or anti- $\alpha$ -GlcNAc WTA monoclonal antibody, or isotype control antibody, anti-cytomegalovirus glycoprotein-D (gD) to USA300 isolated from infected kidneys ( $n=3$ ). MFI, mean fluorescence intensity. **d**, Binding of anti-GlcNAc WTA antibodies (red) or isotype control (grey) to protein-A-deficient USA300 lacking *tarM* or *tarS* ( $n=3$ ). WT, wild type. **e**, Crystal structure of anti- $\beta$ -GlcNAc WTA Fab bound to a synthetic minimal  $\beta$ -WTA unit.

bacterial RNA polymerase are conserved between rifampicin and the rifalogue<sup>25</sup>. Consistent with this, the minimum inhibitory concentration (MIC) for the rifalogue is identical to rifampicin ( $4 \times 10^{-9}$  M), as measured in a conventional growth inhibition assay using planktonic bacteria (Fig. 2f). The rifalogue, but not rifampicin, resulted in a more than 1,000-fold decrease in the number of viable, but non-replicating, bacteria after overnight incubation in minimal PBS buffer (Fig. 2g). The rifalogue was also able to kill classically defined persister cells, bacteria that presumably enter a dormant state to survive antibiotic treatment (for example, ciprofloxacin) of growing cultures<sup>26</sup> (Fig. 2h), while treatment with rifampicin had no effect on recovery of persister cells, in agreement with previous observations<sup>26</sup>. Additionally, rifalogue, but not rifampicin, was able to kill non-dividing MRSA sequestered inside murine macrophages, in part due to its capacity to accumulate rapidly in mammalian cells (Extended Data Fig. 3). As predicted, the rifalogue had no ability to kill extracellular bacteria when linked to the anti- $\beta$ -WTA monoclonal antibody in the AAC format, but was fully bactericidal after the AAC was treated with cathepsin B to release the active antibiotic (Fig. 2i).

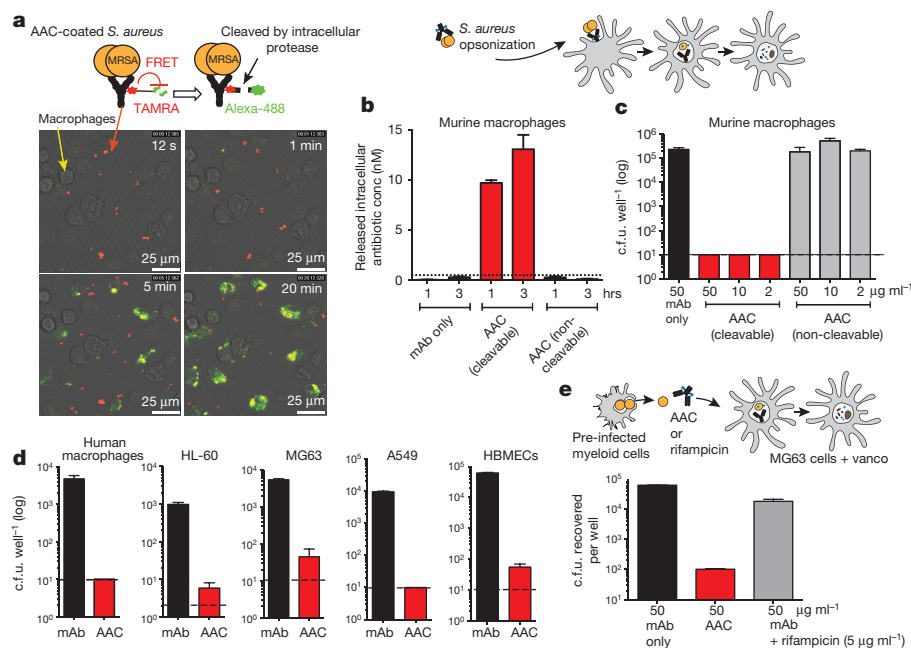
To confirm that the linker used in the AAC is cleaved after internalization of the AAC inside cells, we examined cleavage of the linker with a fluorescence resonance energy transfer (FRET)-based probe consisting of the same anti-MRSA antibody conjugated to two dye molecules that were separated by the same linker used in the AAC<sup>27</sup>. When the

Antibody light chain (pink) and heavy chain (blue) are shown. **f**, MIC determination for rifampicin and rifalogue on USA300 ( $n=5$ ). **g**, Survival of stationary phase USA300 incubated with  $1 \times 10^{-6}$  M rifampicin or rifalogue ( $n=4$ ). **h**, USA300 bacteria were incubated without antibiotic (black) or with  $3 \mu\text{g ml}^{-1}$  ciprofloxacin (Cipro; green, red and grey).  $1 \mu\text{g ml}^{-1}$  of rifalogue (red) or rifampicin (grey) was added as indicated ( $n=3$ ). **i**, Intact AAC does not kill planktonic bacteria but does after pre-treatment with cathepsin-B ( $n=3$ ). **g–i**, Error bars show s.d. for triplicate samples ( $n$  = biological repeats).

fluorescent molecules are covalently associated, fluorescence at 488 nm is quenched; upon cleavage of the linker, quenching is lost. Video microscopy demonstrated that the linker was cleaved within minutes of uptake of the MRSA opsonized with the FRET conjugate (Fig. 3a). We confirmed by mass spectrometry analysis that free rifalogue was released inside macrophages after uptake of MRSA opsonized with active AACs, but not when opsonized with a non-cleavable version of the AAC, which was prepared by replacement of the natural amino acid L-citrulline with D-citrulline at the P1 position, making the linker uncleavable by cathepsins (Fig. 3b).

### AACs eradicate intracellular *S. aureus* infections

When the AAC was used to opsonize MRSA, it readily killed bacteria inside the macrophages, whereas the same bacteria opsonized with non-cleavable  $\beta$ -WTA AAC or monoclonal antibody alone survived (Fig. 3c). AAC-opsonized MRSA were killed inside every cell type tested, including human macrophages, endothelial and epithelial cell lines (Fig. 3d). These data demonstrate that the AAC releases active antibiotic and kills *S. aureus* only after internalization inside host cells, providing a unique tool to assess the importance of intracellular *S. aureus* in various settings. To determine whether specifically targeting intracellular *S. aureus* could prevent cell-to-cell transfer of bacteria, infected peritoneal cells were incubated with MG63 osteoblast cells in the presence of vancomycin (Fig. 3e). Addition of AAC to the



**Figure 3 | AAC linker is cleaved after internalization of bacteria.**

**a**, Live cell imaging monitoring cleavage of AAC linker in macrophages with FRET-based antibody conjugate (representative of three fields). TAMRA, tetramethylrhodamine. **b**, Mass spectrometric quantification of released antibiotic inside macrophages from AAC made with standard (cleavable) or non-cleavable linker. mAb, monoclonal antibody. **c**, **d**, USA300, opsonized with antibody, AAC, or non-cleavable AAC,

enumerated 48 h after incubation with indicated cells lines. HBMECs, human brain microvascular endothelial cells. Horizontal line indicates the limit of detection. **e**, Intracellular USA300 as in Fig. 1a were added to a monolayer of MG63 with antibody, AAC, or a mixture of antibody plus rifampicin in media containing vancomycin (vanco). Surviving bacteria were enumerated 24 h later. **b–e**, Error bars represent means  $\pm$  s.d. from triplicate wells. Data are representative of 2–4 biological replicates.

culture media resulted in a significant reduction in the number of viable intracellular bacteria recovered 1 day after infection. Similar treatment with the unconjugated anti-MRSA antibody plus a tenfold higher molar concentration of rifampicin, purported to be one of the best-known antibiotics for treatment of intracellular bacteria<sup>28</sup>, had very marginal efficacy. To confirm that targeting intracellular *S. aureus* could also prevent transfer of infection from intracellular reservoirs *in vivo*, mice were infected by intravenous injection with equivalent doses of planktonic or intracellular MRSA and then treated with PBS or vancomycin, with or without a single dose of AAC. In the presence of vancomycin only intracellular bacteria were able to efficiently colonize the brain (Extended Data Fig. 1); however, treatment with a single dose of AAC effectively eliminated the bacteria that had escaped vancomycin treatment (Extended Data Fig. 4)

### AAC is superior to vancomycin *in vivo*

To determine the role of intracellular infection in bacteraemia, AACs were tested in an intravenous infection model. First, we tested the efficacy of unconjugated anti-MRSA antibodies by treating mice with either anti- $\beta$ -WTA antibodies, anti- $\alpha$ -WTA antibodies or intravenous immunoglobulin (IGIV), a pooled immunoglobulin preparation from ~10,000 humans, 1 h before intravenous infection with wild-type USA300 MRSA. *S. aureus* is a common colonizer of human skin and mucosal surfaces and preliminary analysis of multiple sources of human serum, including IGIV, demonstrated that human serum contains approximately 300  $\mu$ g ml<sup>-1</sup> of anti-*S. aureus* antibodies, of which ~70% are directed towards the GlcNAc modifications of WTA (Extended Data Fig. 5). Even in this prophylactic treatment setting, unconjugated antibodies did not have any efficacy in preventing infection, whereas treatment with vancomycin eliminated detectable bacteria (Fig. 4a). However, the efficacy of vancomycin was limited when treatment was initiated several hours after infection was established (Fig. 4b). In at least one model of bacteraemia, 95% of the bacteria in blood are found to be associated with neutrophils within 15 min (ref. 6), suggesting that failure of vancomycin in this model could be

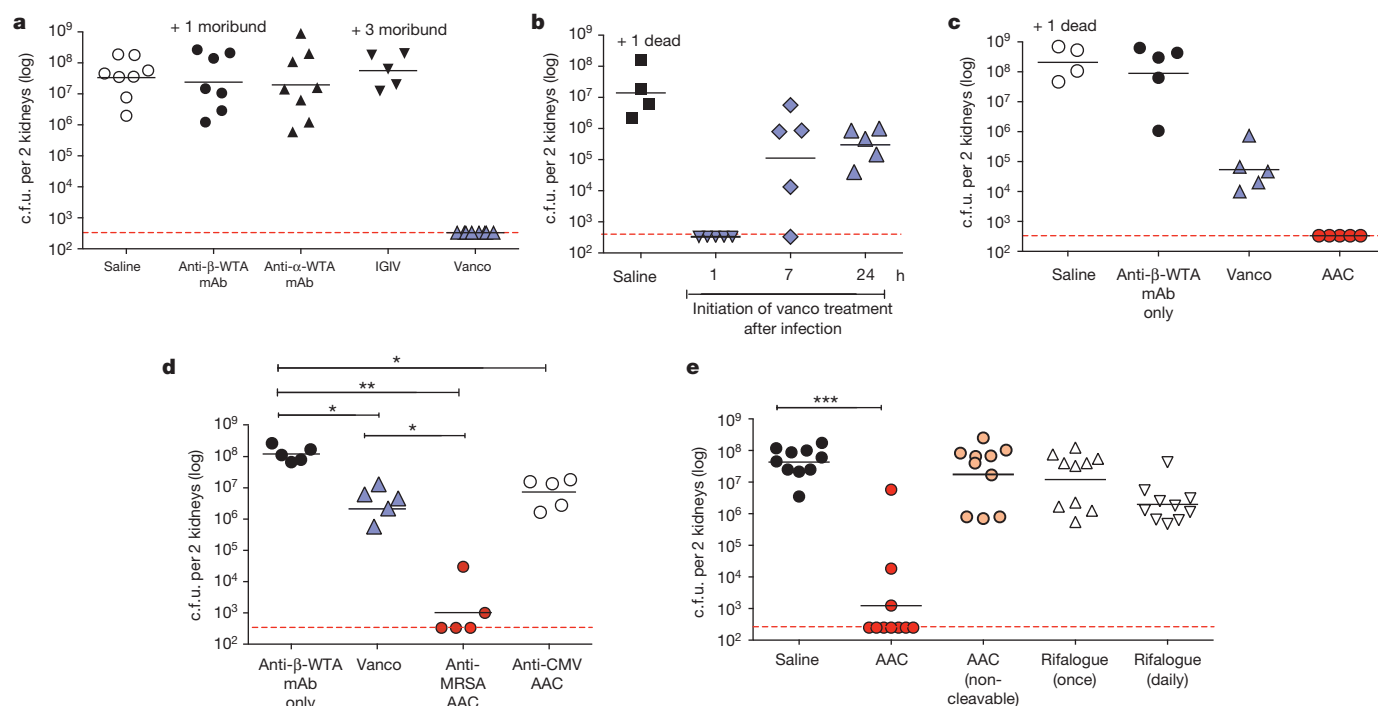
due to antibiotic escape through survival inside host cells. To test this directly, mice were treated with a single dose of AAC 24 h after infection. A single dose of AAC was efficacious and proved superior to twice daily vancomycin treatment initiated at the same time (Fig. 4c and Extended Data Fig. 6a).

Mouse serum has no appreciable anti-*S. aureus* antibody. To determine whether endogenous anti-WTA antibodies found in normal human serum might compete for binding with the AAC, SCID mice were reconstituted with physiological levels of human IgG by daily treatment with IGIV. The resulting SCID-huIgG mice had sustained levels of at least 10 mg ml<sup>-1</sup> of human IgG in the serum and were equally susceptible to infection with MRSA compared to untreated controls (data not shown). Despite the presence of potentially competing antibodies, a single dose of anti-MRSA AAC administered at 24 h after infection with MRSA was still more efficacious than vancomycin (Fig. 4d). An AAC made with an irrelevant antibody (anti-cytomegalovirus glycoprotein D antibody)—which could still bind to protein A on MRSA—had much less efficacy. AACs made with anti- $\beta$ -GlcNAc WTA antibodies appeared to be more efficacious than those made with anti- $\alpha$ -GlcNAc WTA antibodies (Extended Data Fig. 6b), suggesting that the extent of antibody binding determines antibiotic delivery. Treatment with the AAC was more effective than treatment with an equivalent dose or repeated dosing of unconjugated rifampicin, and release of the antibiotic from the AAC was essential for efficacy, as an AAC generated with the non-cleavable linker was not efficacious *in vivo* (Fig. 4e).

### Discussion

Treatment of patients with invasive MRSA infections with conventional antibiotics results in failure rates of up to 50% (refs 29–32), in most cases without measurable outgrowth of antibiotic-resistant strains. Survival of antibiotic-susceptible subpopulations of *S. aureus* in the presence of antibiotic treatment is well documented and is probably due to multiple factors, including survival of persister bacteria that are not actively dividing and poor antibiotic exposure of





**Figure 4 | AAC is a more effective treatment than vancomycin after intravenous infection.** **a**, Wild-type (WT) mice ( $n = 8$  per group) were treated with  $50 \text{ mg kg}^{-1}$  of the indicated anti-MRSA antibodies 1 h before MRSA infection or twice daily with  $110 \text{ mg kg}^{-1}$  vancomycin (Vanco). **b**, Treatment of wild-type mice ( $n = 5$  per group) with  $110 \text{ mg kg}^{-1}$  vancomycin (twice daily) was initiated either at 1 h, 7 h or 24 h after infection. **c**, Wild-type mice ( $n = 5$  per group) were treated with saline, anti- $\beta$ -WTA antibody used in the AAC (monoclonal antibody (mAb)), vancomycin (twice daily), or anti-MRSA AAC (a single dose of  $50 \text{ mg kg}^{-1}$ ) starting 24 h after infection. **d**, **e**, SCID mice (**d**, 5 mice per

group; **e**, 10 mice per group) were injected with human IgG to achieve a concentration of  $10 \text{ mg ml}^{-1}$ , then infected with MRSA. Treatment as indicated was begun 24 h after infection. Vancomycin,  $110 \text{ mg kg}^{-1}$  twice daily; AAC,  $50 \text{ mg kg}^{-1}$  a single dose on day 1; rifalogue,  $0.5 \text{ mg kg}^{-1}$  either once on day 1, or once daily (days 1–3). **a–e**, Bacteria in kidneys were determined 4 days after infection. Schematic of each experimental design is shown in Extended Data Fig. 7. Each point shows data from a single animal. Bars show geometric mean. Dashed lines indicate limit of detection. Mann–Whitney  $U$ -test. \* $P < 0.05$ ; \*\* $P < 0.01$ ; \*\*\* $P < 0.005$ .

bacteria in protected niches<sup>33–35</sup>. However, it has been difficult to determine conclusively what role intracellular *S. aureus* plays in disease pathogenesis and antibiotic failure. To address directly the extent to which this ability to survive and replicate inside cells contributes to the failure of conventional antibiotics, we developed a novel therapeutic in the form of an AAC that is specifically designed to be activated only within host cells. We found that a single dose of AAC treatment is effective in a murine model of bacteraemia, in which vancomycin, the current standard of care for MRSA infection, failed (Fig. 4). This suggests that ablation of the intracellular pool of pathogens is key for clinical success in the treatment of potential fatal MRSA infection.

In designing the AAC to test this hypothesis, we used an antibiotic that kills intracellular bacteria. This rifalogue antibiotic also ablates non-replicating bacteria, and antibiotic-resistant persister cells. We posit that this characteristic, along with its ability to accumulate within the intracellular milieu, is essential for AAC success. When the closely related rifampicin, which does not have these characteristics, was conjugated to the same anti-WTA antibody it did not kill intracellular bacteria and exhibited poor efficacy *in vivo* (Extended Data Fig. 8). These studies show that arming antibodies with unique bactericidal antibiotics can result in promising new therapies. Many potent antibiotic-like compounds with the desired characteristics to kill difficult-to-treat intracellular pathogens fail in clinical practice owing to poor pharmacokinetic properties or undesired host toxicity. The use of an AAC might be able to overcome these problems, making it a novel platform for delivering potent antibacterial compounds that may not have a suitable profile as unconjugated drugs. Antibiotic failure has been associated with the ability of a variety of bacterial pathogens to survive within host phagocytic cells<sup>36</sup>. The

extent to which other common human pathogens beyond *S. aureus* depend on this intracellular protected niche merits investigation; the AAC platform promises to enhance antibiotic efficacy against these infectious diseases.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 15 December 2014; accepted 6 October 2015.**

**Published online 4 November 2015.**

- Diekema, D. J. *et al.* Survey of infections due to *Staphylococcus* species: frequency of occurrence and antimicrobial susceptibility of isolates collected in the United States, Canada, Latin America, Europe, and the Western Pacific region for the SENTRY Antimicrobial Surveillance Program, 1997–1999. *Clin. Infect. Dis.* **32** (suppl. 2), S114–S132 (2001).
- Lowy, F. D. *Staphylococcus aureus* infections. *N. Engl. J. Med.* **339**, 520–532 (1998).
- Boucher, H. W. *et al.* Bad bugs, no drugs: no ESCAPE! An update from the Infectious Diseases Society of America. *Clin. Infect. Dis.* **48**, 1–12 (2009).
- Nannini, E., Murray, B. E. & Arias, C. A. Resistance or decreased susceptibility to glycopeptides, daptomycin, and linezolid in methicillin-resistant *Staphylococcus aureus*. *Curr. Opin. Pharmacol.* **10**, 516–521 (2010).
- Thwaites, G. E. & Gant, V. Are bloodstream leukocytes Trojan Horses for the metastasis of *Staphylococcus aureus*? *Nature Rev. Microbiol.* **9**, 215–222 (2011).
- Rogers, D. E. & Tompsett, R. The survival of staphylococci within human leukocytes. *J. Exp. Med.* **95**, 209–230 (1952).
- Gresham, H. D. *et al.* Survival of *Staphylococcus aureus* inside neutrophils contributes to infection. *J. Immunol.* **164**, 3713–3722 (2000).
- Kapral, F. A. & Shayegani, M. G. Intracellular survival of staphylococci. *J. Exp. Med.* **110**, 123–138 (1959).
- Anwar, S., Prince, L. R., Foster, S. J., Whyte, M. K. & Sabroe, I. The rise and rise of *Staphylococcus aureus*: laughing in the face of granulocytes. *Clin. Exp. Immunol.* **157**, 216–224 (2009).
- Fraunholz, M. & Sinha, B. Intracellular *Staphylococcus aureus*: live-in and let die. *Front. Cell. Infect. Microbiol.* **2**, 43 (2012).

11. Garzoni, C. & Kelley, W. L. Return of the Trojan horse: intracellular phenotype switching and immune evasion by *Staphylococcus aureus*. *EMBO Mol. Med.* **3**, 115–117 (2011).
12. Rogers, D. E. Studies on bacteremia. I. Mechanisms relating to the persistence of bacteremia in rabbits following the intravenous injection of staphylococci. *J. Exp. Med.* **103**, 713–742 (1956).
13. Velasco, E. *et al.* Comparative study of clinical characteristics of neutropenic and non-neutropenic adult cancer patients with bloodstream infections. *Eur. J. Clin. Microbiol. Infect. Dis.* **25**, 1–7 (2006).
14. Venditti, M. *et al.* *Staphylococcus aureus* bacteremia in patients with hematologic malignancies: a retrospective case-control study. *Haematologica* **88**, 923–930 (2003).
15. Bosse, M. J., Gruber, H. E. & Ramp, W. K. Internalization of bacteria by osteoblasts in a patient with recurrent, long-term osteomyelitis. A case report. *J. Bone Joint Surg. Am.* **87**, 1343–1347 (2005).
16. Clement, S. *et al.* Evidence of an intracellular reservoir in the nasal mucosa of patients with recurrent *Staphylococcus aureus* rhinosinusitis. *J. Infect. Dis.* **192**, 1023–1028 (2005).
17. Jarry, T. M., Memmi, G. & Cheung, A. L. The expression of  $\alpha$ -haemolysin is required for *Staphylococcus aureus* phagosomal escape after internalization in CFT-1 cells. *Cell. Microbiol.* **10**, 1801–1814 (2008).
18. Que, Y. A. *et al.* Fibrinogen and fibronectin binding cooperate for valve infection and invasion in *Staphylococcus aureus* experimental endocarditis. *J. Exp. Med.* **201**, 1627–1635 (2005).
19. Greenlee-Wacker, M. C. *et al.* Phagocytosis of *Staphylococcus aureus* by human neutrophils prevents macrophage efferocytosis and induces programmed necrosis. *J. Immunol.* **192**, 4709–4717 (2014).
20. Kobayashi, S. D. *et al.* Rapid neutrophil destruction following phagocytosis of *Staphylococcus aureus*. *J. Innate Immun.* **2**, 560–575 (2010).
21. Barcia-Macay, M., Seral, C., Mingeot-Leclercq, M. P., Tulkens, P. M. & Van Bambeke, F. Pharmacodynamic evaluation of the intracellular activities of antibiotics against *Staphylococcus aureus* in a model of THP-1 macrophages. *Antimicrob. Agents Chemother.* **50**, 841–851 (2006).
22. Sandberg, A., Hessler, J. H., Skov, R. L., Blom, J. & Frimodt-Møller, N. Intracellular activity of antibiotics against *Staphylococcus aureus* in a mouse peritonitis model. *Antimicrob. Agents Chemother.* **53**, 1874–1883 (2009).
23. Dubowchik, G. M. *et al.* Cathepsin B-labile dipeptide linkers for lysosomal release of doxorubicin from internalizing immunoconjugates: model studies of enzymatic drug release and antigen-specific *in vitro* anticancer activity. *Bioconjug. Chem.* **13**, 855–869 (2002).
24. Winstel, V., Xia, G. & Peschel, A. Pathways and roles of wall teichoic acid glycosylation in *Staphylococcus aureus*. *Int. J. Med. Microbiol.* **304**, 215–221 (2014).
25. Campbell, E. A. *et al.* Structural mechanism for rifampicin inhibition of bacterial RNA polymerase. *Cell* **104**, 901–912 (2001).
26. Conlon, B. P. *et al.* Activated ClpP kills persisters and eradicates a chronic biofilm infection. *Nature* **503**, 365–370 (2013).
27. Fischer, R., Hufnagel, H. & Brock, R. A doubly labeled penetratin analogue as a ratiometric sensor for intracellular proteolytic stability. *Bioconjug. Chem.* **21**, 64–73 (2010).
28. Nielsen, S. L. & Black, F. T. Extracellular and intracellular killing in neutrophil granulocytes of *Staphylococcus aureus* with rifampicin in combination with dicloxacillin or fusidic acid. *J. Antimicrob. Chemother.* **43**, 407–410 (1999).
29. Kullar, R., Davis, S. L., Levine, D. P. & Rybak, M. J. Impact of vancomycin exposure on outcomes in patients with methicillin-resistant *Staphylococcus aureus* bacteremia: support for consensus guidelines suggested targets. *Clin. Infect. Dis.* **52**, 975–981 (2011).
30. Fowler, V. G. Jr *et al.* Daptomycin versus standard therapy for bacteremia and endocarditis caused by *Staphylococcus aureus*. *N. Engl. J. Med.* **355**, 653–665 (2006).
31. Yoon, Y. K., Kim, J. Y., Park, D. W., Sohn, J. W. & Kim, M. J. Predictors of persistent methicillin-resistant *Staphylococcus aureus* bacteraemia in patients treated with vancomycin. *J. Antimicrob. Chemother.* **65**, 1015–1018 (2010).
32. Johnson, L. B., Almoujahed, M. O., Ilg, K., Maalood, L. & Khatib, R. *Staphylococcus aureus* bacteremia: compliance with standard treatment, long-term outcome and predictors of relapse. *Scand. J. Infect. Dis.* **35**, 782–789 (2003).
33. Levin, B. R. Noninherited resistance to antibiotics. *Science* **305**, 1578–1579 (2004).
34. Grant, S. S., Kaufmann, B. B., Chand, N. S., Haseley, N. & Hung, D. T. Eradication of bacterial persisters with antibiotic-generated hydroxyl radicals. *Proc. Natl Acad. Sci. USA* **109**, 12147–12152 (2012).
35. Lewis, K. Persister cells. *Annu. Rev. Microbiol.* **64**, 357–372 (2010).
36. Kaiser, P. *et al.* Cecum lymph node dendritic cells harbor slow-growing bacteria phenotypically tolerant to antibiotic treatment. *PLoS Biol.* **12**, e1001793 (2014).
37. Bryskier, A. Anti-MRSA agents: under investigation, in the exploratory phase and clinically available. *Expert Rev. Anti Infect. Ther.* **3**, 505–553 (2005).

**Acknowledgements** This research used resources of the Advanced Photon Source, a US Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory under contract no. DE-AC02-06CH11357.

**Author Contributions** S.M.L. designed and executed the *in vitro* and *in vivo* analysis of the AAC mechanism of action. T.P., L.S. and J.A.F. designed and synthesized antibiotics and linker drugs. M.X., J.K., S.P. and D.Y. designed and analysed *in vivo* models for intravenous infection. H.R., L.D., M.D. and R.V. designed and conjugated linker antibiotic to antibodies. K.K.K., W.L.H., J.H.M. and S.M. characterized the anti-MRSA antibodies. Y.K., H.H., K.M.L., E.P. and J.C. did mass spectrometry analysis of the rifalogs during *in vitro* efficacy studies. P.L. and R.F. performed X-ray crystallography of anti- $\beta$ -WTA monoclonal antibody. J.P.L. designed the synthesis of  $\beta$ -phospho-ribitol. B.-C.L. and C.C. characterized FRET constructs and helped with video microscopy. E.L. determined the number of antibody-binding sites on MRSA. M.S., K.K. and P.S.A. isolated anti-MRSA antibodies from patients. M.W.T. contributed to bacterial genetics and data analysis. E.J.B. and S.M. initiated the project and S.M. led the project. S.M.L., E.J.B. and S.M. composed the paper with input from all authors.

**Additional Information** The structure of the anti- $\beta$ -WTA Fab bound to the synthetic WTA fragment ( $\beta$ -GlcNAc anomer) has been deposited in the Protein Data Bank under accession number 5D6C. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.J.B. ([brown.eric@gene.com](mailto:brown.eric@gene.com)) or S.M. ([sanj@gene.com](mailto:sanj@gene.com)).



## METHODS

**Ethics statement.** All animal procedures were conducted under a protocol (#08–1990) approved by the Genentech Institutional Animal Care and Use Committee in an Association for Assessment and Accreditation of Laboratory Animal Care (AAALAC)-accredited facility in accordance with the Guide for the Care and Use of Laboratory Animals and applicable laws and regulations. For cloning of antibodies from human B cells, informed written consent was obtained from all donors and was provided in accordance with the Declaration of Helsinki. Approval was obtained from the health research ethics committee of Denmark through the regional committee for the Capital Region of Denmark.

**Bacterial strains.** All *in vivo* experiments were done with MRSA-USA300 NRS384 obtained from NARSA (<https://www.beiresources.org>) unless noted otherwise. The generation of protein-A-deficient strain  $\Delta mcr$  USA300 NRS384, as well as protein-A-deficient USA300 lacking *tarM* or *tarS* has been described previously<sup>38,39</sup>. The protein-A-deficient strains were used only in some *in vitro* experiments to determine antibody specificity.

**MIC determinations for extracellular bacteria.** The MIC for extracellular bacteria was determined by preparing serial twofold dilutions of the antibiotic in tryptic soy broth. Dilutions of the antibiotic were made in quadruplicate in 96-well culture dishes. MRSA (NRS384 strain of USA300) was taken from an exponentially growing culture and diluted to  $1 \times 10^4$  c.f.u. ml<sup>-1</sup>. The bacteria were cultured in the presence of antibiotic for 18–24 h with shaking at 37°C and bacterial growth was determined by reading the optical density (OD) at 630 nm. The MIC was determined to be the dose of antibiotic that inhibited bacterial growth by >90%.

**MIC determinations for intracellular bacteria.** Intracellular MIC was determined on bacteria that were sequestered inside mouse peritoneal macrophages (see later for generation of murine peritoneal macrophages). Macrophages were plated at a density of  $4 \times 10^5$  cells ml<sup>-1</sup> and infected with MRSA at a ratio of 10–20 bacteria per macrophage. Macrophage cultures were maintained in growth media supplemented with  $50 \mu\text{g ml}^{-1}$  of gentamycin to inhibit the growth of extracellular bacteria and test antibiotics were added to the growth media 1 day after infection. The survival of intracellular bacteria was assessed 24 h after addition of the antibiotics. Macrophages were lysed with Hanks buffered saline solution supplemented with 0.1% bovine serum albumin (BSA) and 0.1% Triton-X, and serial dilutions of the lysate were made in PBS solution containing 0.05% Tween-20. The number of surviving intracellular bacteria was determined by plating on tryptic soy agar plates with 5% defibrinated sheep blood.

***In vivo* transfer of infection model.** USA300 stocks were prepared for infection from actively growing cultures in tryptic soy broth. Bacteria were washed three times in PBS and aliquots were frozen at –80°C in PBS 25% glycerol.

**Intracellular bacteria infections.** Seven-week-old female A/J mice (Stock 000646) were obtained from Jackson Labs and infected by peritoneal injection with  $5 \times 10^7$  c.f.u. of USA300. Mice were killed 1 day after infection and the peritoneum was flushed with 5 ml of cold PBS. Peritoneal washes were centrifuged for 5 min at 1,000 r.p.m. at 4°C in a table-top centrifuge. The cell pellet containing peritoneal cells was collected and cells were treated with  $50 \mu\text{g ml}^{-1}$  of lysostaphin (Cell Sciences, CRL 309C) for 20 min at 37°C to kill contaminating extracellular bacteria. Peritoneal cells were washed three times in ice-cold PBS to remove the lysostaphin. Peritoneal cells from donor mice were pooled, and recipient mice were injected with cells derived from five donors per each recipient by intravenous injection into the tail vein. To determine the number of live intracellular colony-forming units, a sample of the peritoneal cells were lysed in HB (Hanks balanced salt solution supplemented with 10 mM HEPES and 0.1% BSA) with 0.1% Triton-X, and serial dilutions of the lysate were made in PBS with 0.05% Tween-20.

**Free bacteria infections.** A/J mice were infected with various doses of free bacteria using a fresh aliquot of the glycerol stocks used for the peritoneal injections. Actual infection doses were confirmed by c.f.u. plating. For the data shown in Fig. 1a the actual infection dose for intracellular bacteria was  $1.8 \times 10^6$  c.f.u. per mouse, and the actual infection dose for free bacteria was  $2.9 \times 10^6$  c.f.u. per mouse. Selected mice were treated with a single dose of  $110 \text{ mg kg}^{-1}$  of vancomycin by intravenous injection immediately after infection.

***In vitro* transfer of infection to non-phagocytic cells.** **Generation of MRSA-infected peritoneal cells.** Six-to-eight-week-old female A/J mice (see earlier) were infected with  $1 \times 10^8$  c.f.u. of the NRS384 strain of USA300 by peritoneal injection. The peritoneal wash was harvested 1 day after infection, and the infected peritoneal cells were treated with  $50 \mu\text{g ml}^{-1}$  of lysostaphin diluted in HEPES buffer supplemented with 0.1% BSA (HB buffer) for 20 min at 37°C. Peritoneal cells were then washed twice in ice-cold HB buffer. The peritoneal cells were diluted to  $1 \times 10^6$  cells ml<sup>-1</sup> in RPMI 1640 tissue culture media supplemented with 10 mM HEPES and 10% fetal calf serum, and  $5 \mu\text{g ml}^{-1}$  vancomycin. Free MRSA from the primary infection was stored overnight at 4°C in PBS

solution as a control for extracellular bacteria that were not subject to neutrophil killing.

**Infection of osteoblasts, HBMEC and A549 cells.** MG63 cell line (CRL-1427) and A549 cells (CCL185) were obtained from ATCC and maintained in RPMI 1640 tissue culture media supplemented with 10 mM HEPES and 10% fetal calf serum (RPMI-10). HBMEC cells (catalogue #1000) and ECM media (catalogue #1001) were obtained from ScienceCell Research Labs. The cells were used without further authentication or testing for mycoplasma contamination. Cells were plated in 24-well tissue culture plates and cultured to obtain a confluent layer. On the day of the experiment, the cells were washed once in RPMI (without supplements). MRSA or infected peritoneal cells were diluted in complete RPMI-10 and vancomycin was added at  $5 \mu\text{g ml}^{-1}$  immediately before infection. Peritoneal cells were added to the osteoblasts at  $1 \times 10^6$  peritoneal cells per ml. A sample of the cells was lysed with 0.1% Triton-X to determine the actual concentration of live intracellular bacteria at the time of infection. The actual titre for all infections was determined by plating serial dilutions of the bacteria on tryptic soy agar with 5% defibrinated sheep blood.

**Generation of the anti-*S. aureus* antibodies.** The human IgG antibodies against anti- $\beta$ -GlcNAc WTA monoclonal antibody (mAb) and anti- $\alpha$ -GlcNAc WTA mAb were cloned from peripheral B cells from patients after *S. aureus* infection using a monoclonal antibody discovery technology that conserves the cognate pairing of antibody heavy and light chains<sup>40</sup>. Antibodies were expressed by transfection of mammalian cells<sup>41</sup>. Supernatants containing full-length IgG1 antibodies were harvested after 7 days and used to screen for antigen binding by enzyme-linked immunosorbent assay (ELISA). These antibodies were positive for binding to cell wall preparations from USA300. Antibodies were subsequently produced in 200-ml transient transfections and purified with protein A chromatography (MabSelect SuRe, GE Life Sciences) for further testing.

**Synthesis of the linker drug.** Synthesis of the rifalogue linker drug was performed as follows. Protease cleavable linker MC-VC-PAB-OH<sup>23</sup> (1.009 g, 1.762 mmol, 1.000, 1,009 mg) was taken up in *N,N*-dimethylformamide (6 ml, 77 mmol, 44, 5,700 mg). To this was added a solution of thionyl chloride (1.1 equiv., 1.938 mmol, 1.100, 231 mg) in dichloromethane (DCM) (1 ml, 15.44 mmol, 8.765, 1,325 mg) in portions dropwise (half was added over 1 h, stirred for 1 h at room temperature, then the other half was added over another hour). The solution remained a yellow colour. Another 0.6 equiv. of thionyl chloride was added as a solution in 0.5 ml DCM dropwise, carefully. The reaction remained yellow and was stirred sealed overnight at room temperature. The reaction was monitored by liquid chromatography mass spectrometry (LC/MS), indicating 88% conversion to benzyl chloride. Another 0.22 equiv. of thionyl chloride was added dropwise as a solution in 0.3 ml DCM. When the reaction approached 92% benzyl chloride, the reaction was bubbled with N<sub>2</sub>. The concentration was increased from 0.3 M to 0.6 M.

MC-VC-PAB-Cl (0.9 mmol) was cooled to 0°C and rifalogue (dimethyl piperazinebenzoxazinorifamycin<sup>42</sup> (0.75 g, 0.81 mmol, 0.46, 750 mg)) was added. The mixture was diluted with another 1.5 ml of DMF to reach 0.3 M. Stirred open to air for 30 min. *N,N*-diisopropylethylamine (3.5 mmol, 3.5 mmol, 2.0, 460 mg) was added and the reaction stirred overnight open to air. Over the course of 4 days, four additions of 0.2 equiv. *N,N*-diisopropylethylamine base were added while the reaction stirred open to air, until the reaction appeared to stop progressing. The reaction was diluted with DMF and purified on high-performance liquid chromatography (HPLC; 20–60% ACN/FA-H<sub>2</sub>O) in several batches to give MC-VC-PAB-rifalogue (0.38 g, 32% yield) *m/z* = 1,482.8.

The non-cleavable rifalogue linker drug was synthesized using the exact same method, but replacing MC-VC-PAB-OH with MC-V-D-Cit-PAB-OH.

**Conjugation of the linker drug to antibody.** Construction and production of the THIOMAB variant of anti-WTA antibody was done as reported previously<sup>43</sup>. Briefly, a cysteine residue was engineered at the Val 205 position of the anti-WTA light chain to produce its THIOMAB variant. The thio anti-WTA was conjugated to MC-vc-PAB-rifalogue. The antibody was reduced in the presence of 50-fold molar excess dithiothreitol (DTT) overnight. The reducing agent and the cysteine and glutathione blocks were purified away using HiTrap SP-HP column (GE Healthcare). The antibody was re-oxidized in the presence of 15-fold molar excess dehydroascorbic acid (MP Biomedical) for 2.5 h. The formation of interchain disulfide bonds was monitored by LC/MS. A threefold molar excess of the linker drug (MC-VC-PAB-rifalogue) over protein was incubated with the THIOMAB for 1 h. The AAC was purified by filtration through a 0.2  $\mu\text{m}$  SFCA filter (Millipore). Excess-free linker drug was removed by filtration. The conjugate was buffer exchanged into 20 mM histidine acetate pH 5.5/240 mM sucrose by dialysis. The number of conjugated MC-VC-PAB-rifalogue molecules per mAb was quantified by LC/MS analysis. Purity was also assessed by size-exclusion chromatography.

**Mass spectrometric analysis.** LC/MS analysis was performed on a 6530 Accurate-Mass Quadrupole Time-of-Flight (Q-TOF) LC/MS (Agilent Technologies). Samples were chromatographed on a PRLP-S column, 1,000 Å, 8 µm (50 mm × 2.1 mm, Agilent Technologies) heated to 80 °C. A linear gradient from 30–60% B in 4.3 min (solvent A, 0.05% TFA in water; solvent B, 0.04% TFA in acetonitrile) was used and the eluent was directly ionized using the electrospray source. Data were collected and deconvoluted using the Agilent Mass Hunter qualitative analysis software. Before LC/MS analysis, AAC was treated with lysyl endopeptidase (Wako) for 30 min at 1:100 w/w enzyme to antibody ratio, pH 8.0, and 37 °C to produce the Fab and the Fc portion for ease of analysis. The drug-to-antibody ratio (DAR) was calculated using the abundance of Fab and Fab+1 calculated by the MassHunter software.

**Flow cytometry to compare expression of anti-MRSA antibodies.** Analysis of bacteria isolated from infected mice. Balb/c mice were infected with  $1 \times 10^7$  c.f.u. of MRSA (USA300) by intravenous injection and kidneys were harvested on day 3 after infection. Kidneys were homogenized using a GentleMACS dissociator in 5 ml volume per two kidneys using M-Tubes and the program RNA01.01 (Miltenyi Biotec). Homogenization buffer was: PBS plus 0.1% Triton-X-100, 10 µg ml<sup>-1</sup> DNAase (bovine pancreas grade II, Roche) and protease inhibitors (complete protease inhibitor cocktail, Roche 11-836-153001). After homogenization, the samples were incubated at room temperature for 10 min and then diluted with ice-cold PBS and filtered through a 40 µm cell strainer. Tissue homogenates were washed twice in ice-cold PBS and then suspended in a volume of 0.5 ml per two kidneys in HB buffer (Hanks balanced salt solution supplemented with 10 mM HEPES and 0.1% BSA). The cell suspension was filtered again and 25 µl of the bacterial suspension was taken for each staining reaction (Fig. 2c).

**Antibody staining for flow cytometry.** Bacteria ( $1 \times 10^7$  of *in vitro* grown bacteria (Fig. 2d), or 25 µl of tissue homogenate described earlier (Fig. 2c) were suspended in HB buffer and blocked by incubation with 400 µg ml<sup>-1</sup> of mouse IgG (Sigma, 15381) for 1 h. Fluorescently labelled antibodies were added directly to the blocking reaction and incubated at room temperature for an additional 10–20 min. Bacteria were washed three times in HB buffer and then fixed in PBS 2% paraformaldehyde before FACS analysis. Test antibodies (anti-β-WTA, anti-α-WTA or isotype control-anti CMV-gD) were conjugated with Alexa-488 using amine reactive reagents (Invitrogen, succinimidyl-ester of Alexa Fluor 488, NHS-A488). Antibodies in 50 mM sodium phosphate were reacted with a 5–10-fold molar excess of NHS-A488 in the dark for 2–3 h at room temperature. The labelling mixture was applied to a GE Sepharose S200 column equilibrated in PBS to remove excess reactants from the conjugated antibody. The number of A488 molecules per antibody was determined using the ultraviolet method as described by the manufacturer.

For analysis of bacteria in tissue homogenates a non-competing anti-*S. aureus* antibody (rF1 (ref. 38)) was conjugated to Alexa-647 to distinguish *S. aureus* from similar sized particles. Test antibodies were examined at a range of doses from 80 ng ml<sup>-1</sup> to 50 µg ml<sup>-1</sup>. Flow cytometry was performed using a Beckton Dickson FACS ARIA (BD Biosciences) and analysis was performed using FlowJo analysis software (Flow Jo LLC).

**WTA-antibody complex purification and crystallization.** The anti-β-WTA antibody Fab fragment was expressed in *Escherichia coli* and purified on Protein G Sepharose followed by SP sepharose cation exchange and size-exclusion chromatography. Antibody was concentrated to 30 mg ml<sup>-1</sup> in MES buffer (20 mM MES pH 5.5, 150 mM NaCl) and mixed with a 2:1 mol/mol ratio of the WTA analogue (diluted in water) for crystallization trials. Sparse matrix crystallization screening provided initial hits in PEG-8000 based conditions, which were further optimized to provide diffraction quality crystals. Ultimately, data were collected on a crystal grown by the vapour diffusion method in a sitting drop containing 0.5 µl protein and 0.5 µl 0.08 M sodium cacodylate pH 6.5, 0.16 M calcium acetate, 14.4% PEG-8000, and 20% glycerol. Crystals were cryo-protected in mother liquor, flash frozen in liquid nitrogen, and stored for data collection at 100 K.

**Data collection and structure determination.** Data were collected to 1.7 Å at beamline 22ID at the Advanced Photon Source (APS) under cryo-cooled conditions (100 K) at a wavelength of 1.0 Å. Data were reduced using HKL2000 and SCALEPACK in the space group  $P2_12_12_1$ , with unit cell parameters of  $a = 63.7$  Å,  $b = 111.4$  Å,  $c = 158.4$  Å (see Extended Data Table 1 for processing statistics). The structure was solved by sequential molecular replacement searches using Fab constant and variable regions (Protein Data Bank accession 4I77) as individual search models. Iterative rounds of manual model adjustment with COOT followed by simulated annealing, coordinate, and  $b$ -factor refinement with Phenix and BUSTER (Global Phasing) gave a final model with  $R/R_{\text{free}}$  values of 20.6% and 23.7% respectively. Ramachandran statistics calculated by MolProbity indicate that 97.2% of the model residues lie in favoured regions, with 0.5% outliers.

**Generation of a synthetic WTA fragment.** *Synthesis of dibenzyl phosphorochloridate.* A mixture of NCS (3.5 g, 26.6 mmol) was suspended in toluene (80 ml). Then dibenzyl phosphonate (2.0 g, 7.6 mmol) was added. The mixture was stirred at room temperature overnight. The white solid was filtered off and the organic phase was evaporated to give dibenzyl phosphorochloridate (**1**; 2.1 g, 96%) as light yellow oil. <sup>1</sup>H NMR (300 MHz, CDCl<sub>3</sub>, 25 °C) δ 7.36 (s, 10H), 5.20 (m, 4H).

*Synthesis of 4-O-(2-acetamido-3,4,6-tri-O-acetyl-2-deoxy-β-D-glucopyranosyl)-1-O-acetyl-D-ribitol-5-dibenzylphosphate.* A mixture of **2** (described in ref. 44) (500 mg, 0.95 mmol) dissolved in pyridine (12 ml) was cooled to –30 °C and **1** (described ref. 44) (595 mg, 2.0 mmol) was added, stirring for 2 h at –30 °C and warmed to room temperature for 4 h. The mixture was added to H<sub>2</sub>O, and concentrated *in vacuo*. The residue was purified by column chromatography (silica gel: 200 to ~300 mesh; dichloromethane: methanol in a 30:1 as eluent) to give 4-O-(2-acetamido-3,4,6-tri-O-acetyl-2-deoxy-β-D-glucopyranosyl)-1-O-acetyl-D-ribitol-5-dibenzylphosphate (**3**; 190 mg, 24%) as light yellow solid. <sup>1</sup>H NMR (300 MHz, Acetone-d<sub>6</sub>, 25 °C) δ 7.29–7.23 (m, 10H), 7.08 (d, 1H), 5.08 (t, 1H), 4.99–4.78 (m, 6H), 4.31–3.97 (m, 8H), 3.82–3.63 (m, 3H), 1.88 (s, 3H), 1.86 (s, 6H), 1.79 (s, 3H), 1.69 (s, 3H). LC/MS ( $m/z$ ) ES+ 784 [M+H]<sup>+</sup>.

*Synthesis of 4-O-(2-acetamido-3,4,6-tri-O-acetyl-2-deoxy-β-D-glucopyranosyl)-1-O-acetyl-D-ribitol-5-phosphate.* A mixture of **3** (150 mg, 0.19 mmol) dissolved in MeOH (6 ml) was hydrogenated over 10% Pd/C (20 mg) for 2 h at room temperature. Then the mixture was filtered, and the filtrate was evaporated to give 4-O-(2-acetamido-3,4,6-tri-O-acetyl-2-deoxy-β-D-glucopyranosyl)-1-O-acetyl-D-ribitol-5-phosphate (**4**; 100 mg) as light yellow oil. LC/MS ( $m/z$ ) ES+ 604 [M+H]<sup>+</sup>.

*Synthesis of 4-O-(2-acetamido-2-deoxy-β-D-glucopyranosyl)-D-ribitol-5-phosphate (5).* A mixture of **4** (80 mg, 0.16 mmol) dissolved in MeOH (10 ml) was cooled to 5 °C and K<sub>2</sub>CO<sub>3</sub> (30 mg, 0.21 mmol) was added and stirred at 5 °C for 3 h. The reaction was then quenched with 1 N HCl, and concentrated *in vacuo*. The crude product was purified by gel filtration (LH-20, MeOH) to give 4-O-(2-acetamido-2-deoxy-β-D-glucopyranosyl)-D-ribitol-5-phosphate (**5**; 13.3 mg, 23%) as a white solid <sup>1</sup>H NMR (300 MHz, MeOH-d<sub>4</sub>, 25 °C) δ 4.62 (d, 1H), 4.30–4.02 (m, 3H), 3.92–3.32 (m, 9H), 2.03 (s, 3H). LC/MS ( $m/z$ ) ES+ 436 [M+H]<sup>+</sup>.

**Time of kill for free antibiotics on non-replicating bacteria.** *S. aureus* (USA300) was taken from an overnight stationary phase culture, washed once in PBS and suspended at  $1 \times 10^7$  c.f.u. ml<sup>-1</sup> in PBS with no antibiotic or with  $1 \times 10^{-6}$  M antibiotic in a 10 ml volume in 50 ml polypropylene centrifuge tubes. The bacteria were incubated at 37 °C overnight with shaking. At each time point, three 1 ml samples were removed from each culture and centrifuged to collect the bacteria. Bacteria were washed once with PBS to remove the antibiotic and the total number of surviving bacteria was determined by plating serial dilutions of the bacteria on agar plates.

**Killing of persister cells by free antibiotics.** *S. aureus* (USA300) was taken from an overnight stationary phase culture, washed once in tryptic soy broth (TSB) and then adjusted to a final concentration of  $1 \times 10^7$  c.f.u. ml<sup>-1</sup> in a total volume of 10 ml of either TSB or TSB with ciprofloxacin (0.05 mM). Cultures were incubated with shaking at 37 °C for 6 h and then the second antibiotic, either rifampicin (1 µg ml<sup>-1</sup>) or the rifalogue (1 µg ml<sup>-1</sup>) was added. At the indicated times, samples were removed from each culture, washed once with PBS to remove the antibiotic and re-suspended in PBS. The total number of surviving bacteria was determined by plating serial dilutions of the bacteria on agar plates. At the final time point the remainder of each culture was collected and plated.

**Cathepsin release assay for AAC.** To quantify the amount of active antibiotic released from AACs after treatment with cathepsin B, AACs were diluted to 200 µg ml<sup>-1</sup> in cathepsin buffer (20 mM sodium acetate, 1 mM EDTA, 5 mM L-cysteine, pH 5). Cathepsin-B (from bovine spleen, Sigma C7800) was added at 10 µg ml<sup>-1</sup> and the samples were incubated for 1 h at 37 °C. As a control, AACs were incubated in buffer alone. The reaction was stopped by addition of 9 volumes of bacterial growth media, TSB pH 7.4. To estimate the total release of active antibiotic, serial dilutions of the reaction mixture were made in quadruplicate in TSB in 96-well plates and MRSA (USA300) was added to each well at a final density of  $2 \times 10^3$  c.f.u. ml<sup>-1</sup>. The cultures were incubated overnight at 37 °C with shaking and bacterial growth was measured by reading absorbance at 630 nM using a plate reader.

**Synthesis of anti-β-WTA antibody FRET conjugate.** We synthesized and conjugated a maleimide FRET peptide to the anti-β-WTA THIOMAB antibody. We used a FRET pair of tetramethylrhodamine (TAMRA) and fluorescein. The maleimide FRET peptide was synthesized by standard Fmoc solid-phase chemistry using a PS3 peptide synthesizer (Protein Technologies; B.-C.L., M.D. and R.V., manuscript in preparation)<sup>27</sup>. Briefly, 0.1 mmol of Rink amide resin was used to generate C-terminal carboxamide. We used a Fmoc-Lys(Mtt)-OH at the N- and C-terminal residues in order to remove the Mtt group on the resin and carry out additional



side-chain chemistry to attach TAMRA and fluorescein. The sequence of Val-Cit-Leu was added between the FRET pair as a cathepsin-cleavable spacer. The crude maleimide FRET peptide or maleimidocarbonyl-K(TAMRA)-G-V-Cit-L-K (fluorescein) cleaved off from the resin was subjected to further purification by reverse-phase HPLC with a Jupiter 5u C4 column (5  $\mu$ m, 10 mm  $\times$  250 mm; Phenomenex). Our FRET probe allows monitoring not only of the intracellular trafficking of the antibody conjugate, but also the processing of the linker in the phagolysosome. The intact antibody conjugate fluoresces only in red due to the fluorescence resonance energy transfer from the donor. However, upon the substrate cleavage of the FRET peptide in the phagolysosome, the green fluorescence from the donor is expected to appear.

**Video microscopy to detect cleavage of the linker inside macrophages.** Murine peritoneal macrophages were plated on chamber slides (Ibidi, catalogue 80826) in complete media as described for the macrophage intracellular killing assay. USA300 was labelled with Cell Tracker Violet (Invitrogen C10094) at 100  $\mu$ g ml<sup>-1</sup> in PBS 0.1% BSA by incubation for 30 min at 37 °C. The labelled bacteria were opsonized with the anti- $\beta$ -WTA-FRET probe by incubation for 1 h in HB buffer. Macrophages were washed once immediately before addition of the opsonized bacteria, and bacteria were added to cells at 1  $\times$  10<sup>7</sup> bacteria per ml. For non-phagocytosis controls, the macrophages were pre-treated with 60 nM Latrunculin A (Calbiochem) for 30 min before and during phagocytosis. The slides were placed on the microscope immediately after addition of bacteria to the cells and movies were acquired with a Leica SP5 confocal microscope equipped with an environmental chamber with CO<sub>2</sub> and temperature controllers from Ludin. The images were captured every minute for a total time of 30 min using a Plan APO CS  $\times$  40, N.A: 1.25, oil immersion lens, and the 488 nm and 543 nm laser lines to excite Alexa-488 and TAMRA, respectively. Phase images were also recorded using the 543 nm laser line.

**Quantification of released antibiotic inside macrophages.** Primary murine peritoneal macrophages or RAW 264.7 cells (purchased from ATCC) were infected in 24-well tissue culture dishes as described later for the intracellular killing assay with MRSA opsonized with AAC at 100  $\mu$ g ml<sup>-1</sup> in HB. The RAW 264.7 cells were used without further authentication or testing for mycoplasma contamination. After phagocytosis was complete, the cells were washed and 250  $\mu$ l of complete media plus gentamycin was added to wells and the cells were incubated for the indicated time points. At each time point, the supernatant and cellular fractions were collected followed by acetonitrile (ACN) addition to 75% final concentration and incubated for 30 min. Cell and supernatant extracts were lyophilized by evaporation under N<sub>2</sub> (TurboVap; Biotage) and reconstituted in 100  $\mu$ l of 50% ACN, filtered using a 0.45 glass fibre filter plate (Phenomenex) and analysed by LC/MS/MS as follows.

The rifalogue was separated on an Acquity UPLC (Waters Corporation) under gradient elution using a Phenomenex Kinetex XB-C18 column (100  $\text{\AA}$ , 50  $\times$  2.1 mm internal diameter, 2.6  $\mu$ m particle size). The column was maintained at room temperature. The mobile phase was a mixture of 10 mM ammonium acetate in water containing 0.1% formic acid (A) and 90% acetonitrile (B) at a flow rate of 1 ml min<sup>-1</sup>. The rifalogue was eluted with a gradient of 3–98% B over 1 min, followed by 0.8 min at 98% B, then 0.7 min of 3% B to re-equilibrate the column. The injection volume was 10  $\mu$ l.

The Triple Quad 6500 mass spectrometer (Ab Sciex) was operated in a positive ion multiple reaction-monitoring (MRM) mode. The rifalogue precursor (Q1) ion monitored was 927.6  $m/z$  and the product (Q3) ion monitored was 895.2  $m/z$  with collision energy at 27 eV and declustering potential at 191 V. The MS/MS setting parameters were as follows: ion spray voltage, 5,500 V; curtain gas, 40 psi; nebulizer gas (GS1), 35 psi, (GS2), 50 psi; temperature, 600 °C; and dwell time, 150 ms.

Linear calibration curves were obtained for 0.41–100 nM concentration range by spiking rifalogue into cell or supernatant fractions (lacking MRSA or AAC) that were treated similarly to samples. Concentrations of rifalogue were calculated with MultiQuant software (Ab Sciex).

**In vitro intracellular killing assay.** *Non-phagocytic cell types.* MG63 (CRL-1427) and A549 (CCL185) cell lines were obtained from ATCC and maintained in RPMI 1640 tissue culture media supplemented with 10 mM HEPES and 10% fetal calf serum (RPMI-10). HUVEC cells were obtained from Lonza and maintained in EGM endothelial cell complete media (Lonza). HBMEC cells (catalogue #1000) and ECM media (catalogue #1001) were obtained from ScienceCell Research Labs. The cells were used without further authentication or testing for mycoplasma contamination.

**Murine macrophages.** Peritoneal macrophages were isolated from the peritoneum of 6–8-week-old Balb/c mice (Charles River Laboratories). To increase the yield of macrophages, mice were pre-treated by intraperitoneal injection with 1 ml of thioglycolate media (Becton Dickinson). The thioglycolate media was prepared at

a concentration of 4% in water, sterilized by autoclaving, and aged for 20 days to 6 months before use. Peritoneal macrophages were harvested 4 days after treatment with thioglycolate by washing the peritoneal cavity with cold PBS. Macrophages were plated in DMEM supplemented with 10% fetal calf serum, and 10 mM HEPES, without antibiotics, at a density of 4  $\times$  10<sup>5</sup> cells well<sup>-1</sup> in 24-well culture dishes. Macrophages were cultured overnight to permit adherence to the plate.

**Human M2 macrophages.** CD14<sup>+</sup> monocytes were purified from normal human blood using a Monocyte Isolation Kit II (Miltenyi, catalogue 130-091-153) and plated at 1.5  $\times$  10<sup>5</sup> cells cm<sup>-2</sup> on tissue culture dishes pre-coated with fetal calf serum (FCS) and cultured in RPMI 1640 media with 20% FCS plus 100 ng ml<sup>-1</sup> rhM-CSF. Media was refreshed on day 1 and on day 7, the media was changed to 5% serum plus 20 ng ml<sup>-1</sup> IL-4. Macrophages were used 18 h later.

**Assay protocol.** In all experiments bacteria were cultured in TSB. To assess intracellular killing with AACs, USA300 was taken from an exponentially growing culture and washed in HB. AACs or antibodies were diluted in HB (Hanks balanced salt solution supplemented with 10 mM HEPES and 0.1% BSA) and incubated with the bacteria for 1 h to permit antibody binding to the bacteria (opsonization), and the opsonized bacteria were used to infect macrophages at a ratio of 10–20 bacteria per macrophage (4  $\times$  10<sup>6</sup> bacteria in 250  $\mu$ l of HB per well). Macrophages were pre-washed with serum-free DMEM media immediately before infection, and infected by incubation at 37 °C in a humidified tissue culture incubator with 5% CO<sub>2</sub> to permit phagocytosis of the bacteria. After 2 h, the infection mix was removed and replaced with normal growth media (DMEM supplemented with 10% FCS, 10 mM HEPES) and gentamycin was added at 50  $\mu$ g ml<sup>-1</sup> to prevent growth of extracellular bacteria<sup>45</sup>. At the end of the incubation period, the macrophages were washed with serum-free media, and the cells were lysed in HB supplemented with 0.1% Triton-X (lyses the macrophages without damaging the intracellular bacteria). Serial dilutions of the lysate were made in PBS solution supplemented with 0.05% Tween-20 (to disrupt aggregates of bacteria) and the total number of surviving intracellular bacteria was determined by plating on tryptic soy agar with 5% defibrinated sheep blood.

**ELISA for quantification of anti-MRSA antibodies in human serum.** Cell wall preparations (CWPs) were generated from protein-A-deficient *S. aureus* by incubating 40 mg of pelleted bacteria per ml of 10 mM Tris-HCl (pH 7.4) supplemented with 30% raffinose, 100  $\mu$ g ml<sup>-1</sup> of lysostaphin (Cell Sciences), and EDTA-free protease inhibitor cocktail (Roche), for 30 min at 37 °C. The lysates were centrifuged at 11,600g for 5 min, and the supernatants containing cell wall components were collected.

ELISA experiments were performed using standard protocols. Briefly, plates were pre-coated with CWP and then incubated with human IgG preparations: purified human IGIV Immune Globulin (ASD Healthcare), pooled serum from healthy donors or from MRSA patients. The concentrations of anti-staphylococcal IgG present in the serum or purified IgG were calculated by using a calibration curve that was generated with known concentrations of anti-peptidoglycan mAb (4479) against peptidoglycan.

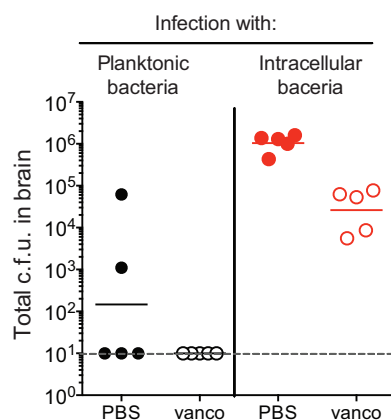
**Intravenous infection model for testing efficacy of the AAC.** Seven-week-old female mice, Balb/c, were obtained from Jackson West, or SCID mice were obtained from Charles River Laboratories. Infections were carried out by intravenous injection into the tail vein. SCID-huIgG model: CB17.SCID mice were reconstituted with IGIV Immune Globulin (ASD Healthcare) using a dosing regimen optimized to achieve constant serum levels of >10 mg ml<sup>-1</sup> of human IgG. IGIV was administered with an initial intravenous dose of 30 mg per mouse followed by a second dose of 15 mg per mouse by intraperitoneal injection after 6 h, and subsequent daily dosings of 15 mg per mouse by intraperitoneal injection for 3 consecutive days. Mice were infected 4 h after the first dose of IGIV with 2  $\times$  10<sup>7</sup> c.f.u. of MRSA diluted in PBS by intravenous injection. The wild-type USA300, protein-A-sufficient strain was used for all *in vivo* experiments. Mice that received vancomycin were treated with twice daily intraperitoneal injections of 110 mg kg<sup>-1</sup> of vancomycin starting between 6 and 24 h after infection for the duration of the study. Experimental therapeutics (AAC, anti-MRSA antibodies or free rifalogue antibiotic) were diluted in PBS and administered with a single intravenous injection 30 min to 24 h after infection. All mice were killed on day 4 after infection, and kidneys were harvested in 5 ml of PBS. The tissue samples were homogenized using a GentleMACS dissociator (Miltenyi Biotec). The total number of bacteria recovered per mouse (two kidneys) was determined by plating serial dilutions of the tissue homogenate in PBS 0.05% Tween on tryptic soy agar with 5% defibrinated sheep blood.

**Statistical analysis.** All experiments were performed on biological replicates. Sample size for each experimental group per condition is reported in appropriate figure legends and Methods. For cell culture experiments, sample size was not predetermined, and all samples were included in the analysis. In animal experiments no statistical methods were used to predetermine sample size ( $n$  = number

of mice per group), and all animals were used for analysis unless the mice died or had to be euthanized when found moribund. These cases are annotated in the figures. The mice were not randomized after infection, and the investigators were not blinded to outcome assessment. When appropriate, statistically significant differences between control and experimental groups were determined using Mann–Whitney tests.

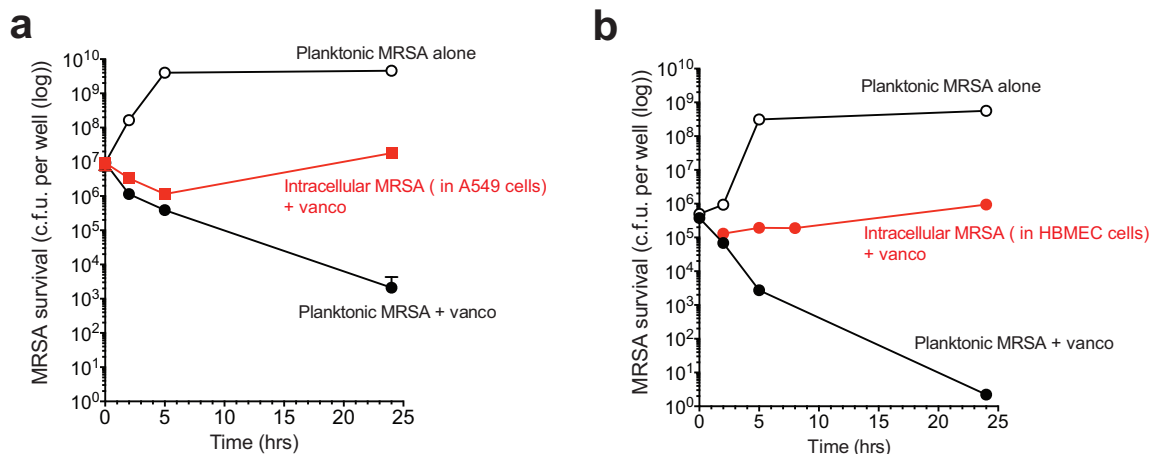
38. Hazenbos, W. L. *et al.* Novel staphylococcal glycosyltransferases SdgA and SdgB mediate immunogenicity and protection of virulence-associated cell wall proteins. *PLoS Pathog.* **9**, e1003653 (2013).
39. Monk, I. R., Shah, I. M., Xu, M., Tan, M. W. & Foster, T. J. Transforming the untransformable: application of direct transformation to manipulate genetically *Staphylococcus aureus* and *Staphylococcus epidermidis*. *MBio* **3**, e00277–11 (2012).
40. Meijer, P. J. *et al.* Isolation of human antibody repertoires with preservation of the natural heavy and light chain pairing. *J. Mol. Biol.* **358**, 764–772 (2006).
41. Meijer, P. J., Nielsen, L. S., Lantto, J. & Jensen, A. Human antibody repertoires. *Methods Mol. Biol.* **525**, 261–277 (2009).
42. Van Duzer, J. *et al.* *Rifamycin Analogs and Uses Thereof* (Activbiotics, 2005).
43. Junutula, J. R. *et al.* Site-specific conjugation of a cytotoxic drug to an antibody improves the therapeutic index. *Nature Biotechnol.* **26**, 925–932 (2008).
44. Boullanger, P., Descotes, G., Flandrois, J. P. & Marmet, D. Synthesis of 4-O-(2-acetamido-2-deoxy- $\beta$ -D-glucopyranosyl)-D-ribitol, antigenic determinant of *Staphylococcus aureus*. *Carbohydr. Res.* **110**, 153–158 (1982).
45. Vaudaux, P. & Waldvogel, F. A. Gentamicin antibacterial activity in the presence of human polymorphonuclear leukocytes. *Antimicrob. Agents Chemother.* **16**, 743–749 (1979).





**Extended Data Figure 1 | Intracellular MRSA can infect the brain even in the presence of vancomycin.** *In vivo* infection of mice shown in

Fig. 1a. Mice ( $n = 5$ ) were infected with equivalent doses of free bacteria or intracellular bacteria and treated with vancomycin at  $110 \text{ mg kg}^{-1}$  10 min after infection and then once per day. Bacterial burden was monitored in the brain 4 days after infection. Bars show geometric mean.

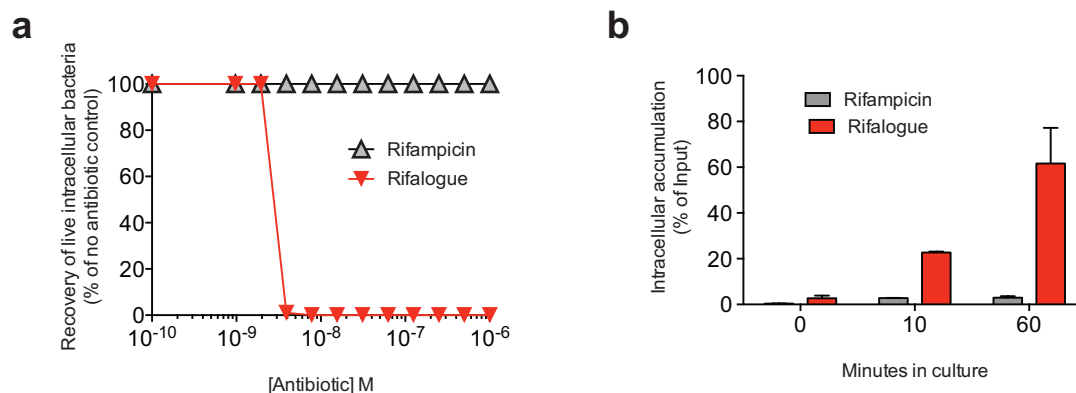


**Extended Data Figure 2 | MRSA is able to grow in the presence of vancomycin when cultured on a monolayer of infectable cells.**

**a, b**, Similar to the set up in Fig. 1d, planktonic MRSA were either seeded in media alone, or in the presence of vancomycin. Intracellular bacteria were generated by infecting a monolayer of either A549 bronchial epithelial cells (**a**) or HBMECs (**b**) in the presence of vancomycin (vanco). In these experiments plates were centrifuged to promote contact of the bacteria with the monolayer to enhance intracellular infection. At each time point, the culture supernatant was collected to recover extracellular

bacteria and adherent cells were lysed to release intracellular bacteria. Extracellular bacteria (planktonic bacteria) grew well in media alone, but were killed by vancomycin. In wells containing a monolayer of mammalian cells (intracellular MRSA + vanco) a fraction of the bacteria were protected from vancomycin during the first 5 h after infection and were able to expand within the intracellular compartment over 24 h. Error bars show s.d. from triplicate wells. Representative of three independent experiments.

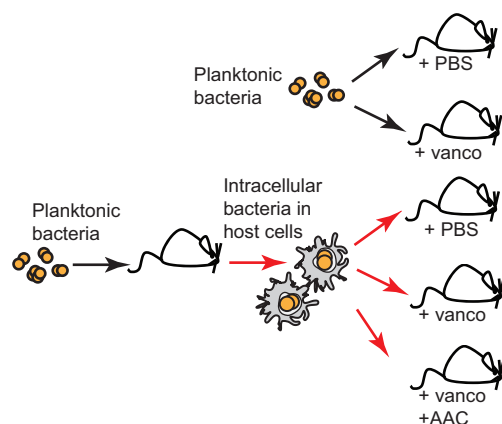




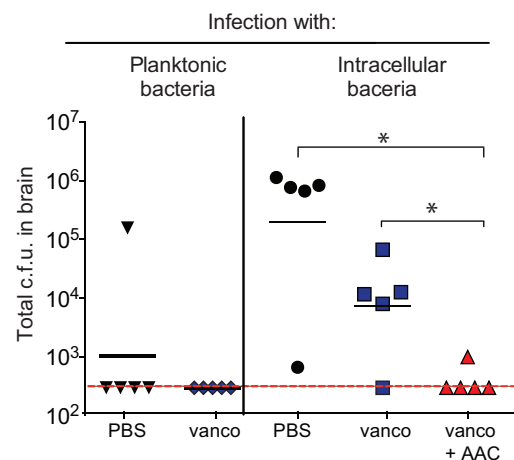
### Extended Data Figure 3 | Rifalogue can also kill intracellular bacteria.

**a**, Determining the intracellular MIC for rifalogue and rifampicin. MRSA was allowed to infect peritoneal macrophages and macrophages were cultured overnight in gentamycin to kill extracellular bacteria. Various doses of rifalogue (red) or rifampicin (grey) were added to the culture medium 1 day after infection and the number of viable intracellular bacteria was determined 24 h later by spotting macrophage lysates onto agar plates. Data shown are representative of more than three independent experiments. **b**, Diffusion of rifalogue versus rifampicin into murine macrophages. Murine peritoneal macrophages were incubated with rifalogue or rifampicin in the culture media. Wells were harvested at 10 and 60 min and the total amount of antibiotic associated with the cells was determined by quantitative mass spectrometry. Results are shown as percentage of input. Error bars show s.d. from triplicate

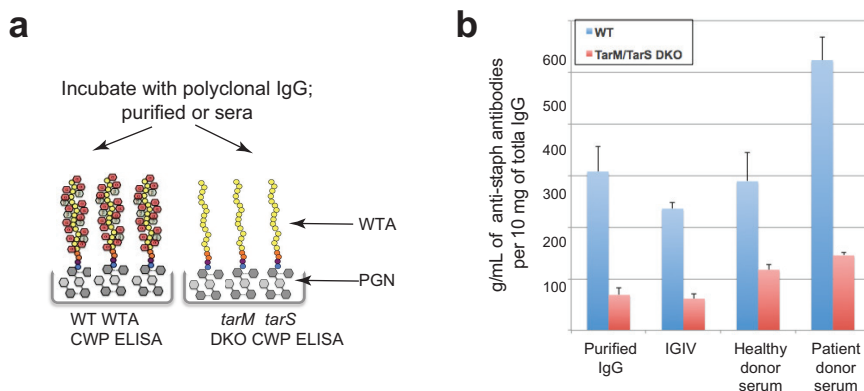
wells. Representative of two experiments. Rifalogue is more lipophilic than rifampicin as a result of its two additional fused aromatic rings, determined by measuring logDs at pH 7, with rifalogue at 3.4, being 100-fold higher than rifampicin (logD 1.3). Additionally, rifalogue has a more basic amine ( $pK_a$  9.7) compared to that found in rifampicin ( $pK_a$  8.2). This balance of lipophilicity and basicity in rifalogue allows it to localize preferentially in lysosomes. It is challenging to develop antibiotics with these properties for systemic administration due to poor pharmacokinetic (PK) properties and toxicity profiles associated with indiscriminate accumulation of these molecules in all host cells. However, appending such antibiotics to an anti-MRSA specific antibody both extends its half-life in the circulation and converts it into an inactive pro-drug whose properties are manifest only after it has been released in phagolysosomes of cells infected with MRSA.



**Extended Data Figure 4 | AAC kills MRSA that survive treatment with vancomycin.** *In vivo* infection of mice as shown in Fig. 1a. Mice were infected with equivalent doses of free bacteria or intracellular bacteria and treated with either saline (PBS) or vancomycin (vanco) at  $110 \text{ mg kg}^{-1}$ , 10 min after infection and then once per day. Selected mice were given



vancomycin as described earlier and also treated with a single dose of AAC at  $50 \text{ mg kg}^{-1}$  10 min after infection. Four days after infection, bacterial burden was monitored in the brain. Each point represents data from a single mouse ( $n = 5$ ). Bars show geometric mean. \* $P < 0.05$ , Mann-Whitney *U*-test.

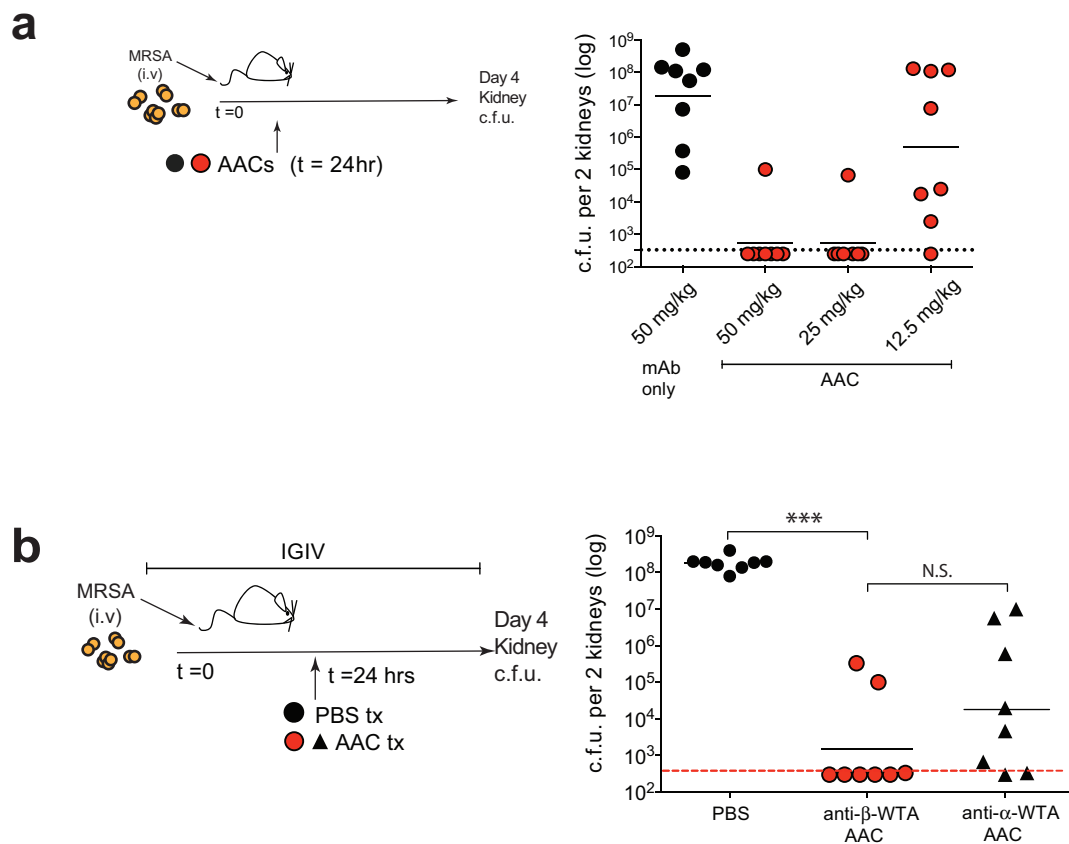


**Extended Data Figure 5 | Human serum contains high levels of anti-*S. aureus* antibodies that can compete with the AAC for binding.**

**a**, To estimate the concentration of antibodies that could potentially compete for binding with the anti- $\beta$ -WTA antibody used in the AAC, human IgG from various sources (normal human serum, serum from MRSA infected patients, purified human IgG (Sigma) or IGIV derived from pooled normal donors) was tested for binding to various bacterial cell wall preparations (CWPs) by enzyme-linked immunosorbent assay (ELISA). CWPs were made from either USA300 (wild type (WT)) or  $\Delta tarM \Delta tarS$  (DKO) USA300; the latter strain is deficient in the WTA-GlcNAc antigen recognized by the anti-WTA antibodies. For these studies protein-A-deficient USA300 background strains were used to

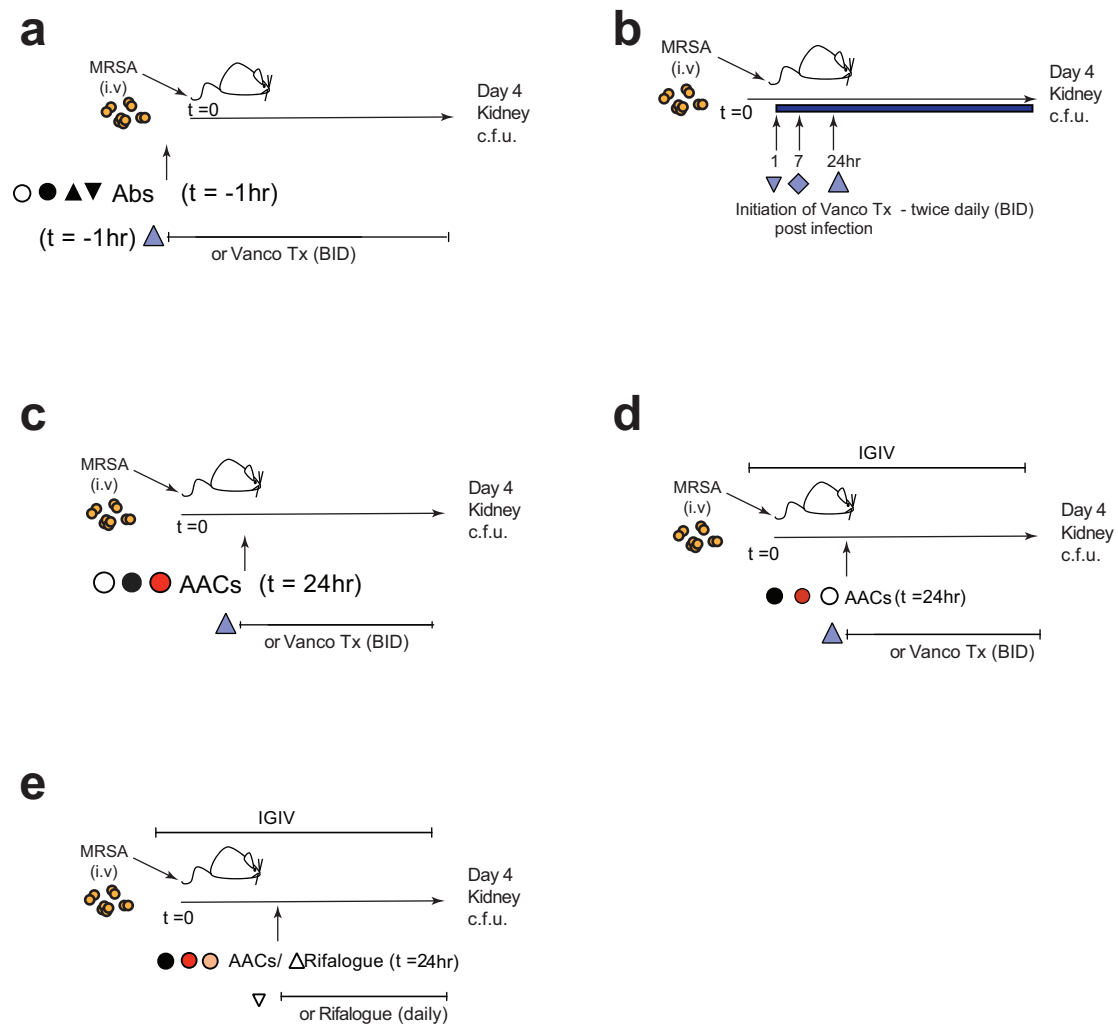
minimize non-specific antibody binding. A standard curve was generated by titrating known amounts of an anti-MRSA antibody directed against peptidoglycan on both cell wall extracts. **b**, Estimated concentration of anti-*S. aureus* antibodies in human serum. The amount of anti-MRSA antibodies in each sample was estimated by comparing the signal obtained for each sample with the standard curve. In the absence of WTA GlcNAc antigens, ~60–70% less serum IgG binding was observed (DKO ELISA; red bars). This indicates the high prevalence of natural antibodies against WTA in adult human serum. Results are reported as  $\mu\text{g ml}^{-1}$  of antibody per 10 mg  $\text{ml}^{-1}$  of total IgG. Error bars show mean  $\pm$  s.d. from triplicate wells. Data are representative of two independent experiments.



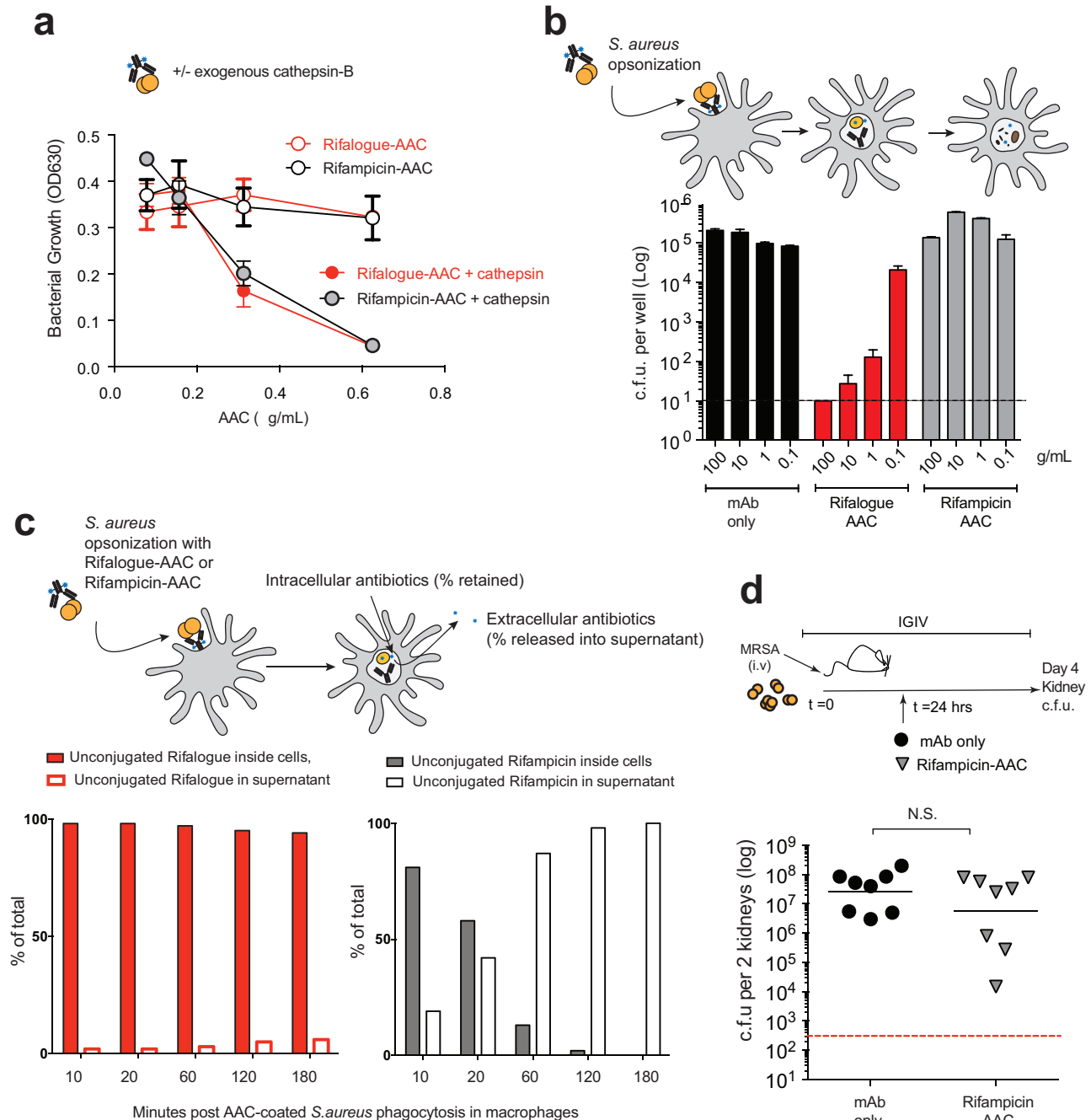


**Extended Data Figure 6 | Optimization of the *in vivo* model of bacteraemia. a,** Titration of AAC in intravenous infection model shown in Fig. 4c ( $n = 8$  mice per group). **b,** Efficacy of AACs specific for  $\beta$ -WTA or  $\alpha$ -WTA in SCID-IGIV model. SCID mice ( $n = 8$  mice per group) were reconstituted with IGIV Immune Globulin using a dosing regimen optimized to achieve constant serum levels of  $>10 \text{ mg ml}^{-1}$  of human

IgG and infected with MRSA. Mice were treated with  $60 \text{ mg kg}^{-1}$  of the indicated AACs in a single intravenous injection 1 day after infection and bacterial burden was monitored in kidneys 4 days after infection. Each point represents data from a single animal. Bars show geometric mean. Mann-Whitney test: \*\*\* $P < 0.005$ , not significant (NS)  $P > 0.05$ .



Extended Data Figure 7 | Schematic of *in vivo* experiments presented in Fig. 4.



**Extended Data Figure 8 | Comparison of anti- $\beta$ -WTA AACs made with rifalogue and rifampicin.** **a**, Rifampicin-AAC and rifalogue-AAC release equivalent amounts of free antibiotic after treatment with exogenous cathepsin-B. Released antibiotics from AACs made with rifalogue (red) and rifampicin (grey) are equally active as they can kill USA300 grown in broth culture. Error bars show s.d. from triplicate wells. **b**, Intracellular killing assay in primary mouse macrophages as shown in Fig. 3c indicates that the rifalogue-AAC, but not rifampicin-AAC is able to kill intracellular *S. aureus*. Error bars show s.d. from triplicate wells. **c**, Greater intracellular retention of unconjugated rifalogue compared with rifampicin after release from AAC inside macrophage cells. MRSA

was opsonized with AACs and incubated with macrophages (RAW 264.7 cells) to permit phagocytosis. The macrophages were washed to remove extracellular bacteria and samples of cell lysates or supernatants were collected in triplicate at indicated time points and the total amount of released antibiotic was determined by quantitative mass spectrometry. Error bars represent means  $\pm$  s.d. from triplicate wells. **a-c**, Representative of two or more independent experiments. **d**, Rifampicin-AAC is not efficacious in the SCID-IGIV intravenous infection model as shown in Fig. 4d. Each point represents data from a single mouse ( $n=8$  mice per group). Bars show geometric mean. Mann-Whitney  $U$ -test: not significant (NS),  $P > 0.05$ .



Extended Data Table 1 | Data collection and refinement statistics for anti- $\beta$ -WTA-WTA complex

	<u>anti-<math>\beta</math>-WTA Fab-</u> <u>WTA complex</u>
<b>Data collection</b>	APS SER-CAT 22ID
Space group	P2 <sub>1</sub> 2 <sub>1</sub> 2 <sub>1</sub>
Cell dimensions	
<i>a</i> , <i>b</i> , <i>c</i> (Å)	63.71, 111.47, 158.41
$\alpha$ , $\beta$ , $\gamma$ (°)	90, 90, 90
Resolution (Å)	50-1.72 (1.78-1.72)
<i>R</i> <sub>sym</sub> or <i>R</i> <sub>merge</sub>	0.056 (0.839)
<i>I</i> / $\sigma I$	29.4 (2.2)
Completeness (%)	99.9 (99.9)
Redundancy	6.0 (5.9)
<b>Refinement</b>	
Resolution (Å)	33.64-1.72
No. <u>reflections</u> (total/test)	120147/6101
<i>R</i> <sub>work</sub> / <i>R</i> <sub>free</sub>	20.6/23.7%
No. <u>atoms</u>	
Protein	6695
WTA	56
Glycerol	12
Calcium	4
Water	694
<i>B</i> -factors	
Protein	35.9
WTA	56.4
Glycerol	49.4
Calcium	38.7
Water	43.3
R.m.s. deviations	
Bond lengths (Å)	0.010
Bond angles (°)	1.13

\*Values in parentheses are for highest-resolution shell.

# Tumour exosome integrins determine organotrophic metastasis

Ayuko Hoshino<sup>1\*</sup>, Bruno Costa-Silva<sup>1\*</sup>, Tang-Long Shen<sup>1,2\*</sup>, Goncalo Rodrigues<sup>1,3</sup>, Ayako Hashimoto<sup>1,4</sup>, Milica Tesic Mark<sup>5</sup>, Henrik Molina<sup>5</sup>, Shinji Kohsaka<sup>6</sup>, Angela Di Giannatale<sup>1</sup>, Sophia Ceder<sup>7</sup>, Swarnima Singh<sup>1</sup>, Caitlin Williams<sup>1</sup>, Nadine Soplop<sup>8</sup>, Kunihiro Uryu<sup>8</sup>, Lindsay Pharmed<sup>9</sup>, Tari King<sup>9</sup>, Linda Bojmar<sup>1,10</sup>, Alexander E. Davies<sup>11</sup>, Yonathan Ararso<sup>1</sup>, Tuo Zhang<sup>12</sup>, Haiying Zhang<sup>1</sup>, Jonathan Hernandez<sup>1,13</sup>, Joshua M. Weiss<sup>1</sup>, Vanessa D. Dumont-Cole<sup>14</sup>, Kimberly Kramer<sup>14</sup>, Leonard H. Wexler<sup>14</sup>, Aru Narendran<sup>15</sup>, Gary K. Schwartz<sup>16</sup>, John H. Healey<sup>17</sup>, Per Sandstrom<sup>10</sup>, Knut Jørgen Labori<sup>18</sup>, Elin H. Kure<sup>19</sup>, Paul M. Grandgenett<sup>20</sup>, Michael A. Hollingsworth<sup>20</sup>, Maria de Sousa<sup>1,3</sup>, Sukhwinder Kaur<sup>21</sup>, Maneesh Jain<sup>21</sup>, Kavita Mallya<sup>21</sup>, Surinder K. Batra<sup>21</sup>, William R. Jarnagin<sup>13</sup>, Mary S. Brady<sup>1,22</sup>, Oystein Fodstad<sup>23,24</sup>, Volkmar Muller<sup>25</sup>, Klaus Pantel<sup>26</sup>, Andy J. Minn<sup>27</sup>, Mina J. Bissell<sup>11</sup>, Benjamin A. Garcia<sup>28</sup>, Yibin Kang<sup>29,30</sup>, Vinagolu K. Rajasekhar<sup>31</sup>, Cyrus M. Ghajar<sup>32</sup>, Irina Matei<sup>1</sup>, Hector Peinado<sup>1,33</sup>, Jacqueline Bromberg<sup>34,35</sup> & David Lyden<sup>1,14</sup>

**Ever since Stephen Paget's 1889 hypothesis, metastatic organotropism has remained one of cancer's greatest mysteries. Here we demonstrate that exosomes from mouse and human lung-, liver- and brain-tropic tumour cells fuse preferentially with resident cells at their predicted destination, namely lung fibroblasts and epithelial cells, liver Kupffer cells and brain endothelial cells. We show that tumour-derived exosomes uptaken by organ-specific cells prepare the pre-metastatic niche. Treatment with exosomes from lung-tropic models redirected the metastasis of bone-tropic tumour cells. Exosome proteomics revealed distinct integrin expression patterns, in which the exosomal integrins  $\alpha_6\beta_4$  and  $\alpha_6\beta_1$  were associated with lung metastasis, while exosomal integrin  $\alpha_5\beta_5$  was linked to liver metastasis. Targeting the integrins  $\alpha_6\beta_4$  and  $\alpha_5\beta_5$  decreased exosome uptake, as well as lung and liver metastasis, respectively. We demonstrate that exosome integrin uptake by resident cells activates Src phosphorylation and pro-inflammatory S100 gene expression. Finally, our clinical data indicate that exosomal integrins could be used to predict organ-specific metastasis.**

Despite Stephen Paget's 126-year-old "seed-and-soil" hypothesis<sup>1</sup>, insufficient progress has been made towards decoding the mechanisms governing organ-specific metastasis. In experimental metastasis assays, Fidler *et al.* demonstrated that cancer cells derived from a certain metastatic site displayed enhanced abilities to metastasize to that specific organ, providing support for Paget's organ-specific metastasis theory<sup>2</sup>. Subsequent studies investigating organ-specific metastasis focused largely on the role of intrinsic cancer cell properties, such as genes and pathways regulating colonization, in directing organotropism<sup>3–8</sup>. Breast cancer cells express chemokine receptors, such as C-X-C motif receptor 4 (CXCR4) and C-C motif receptor 7 (CCR7), which partner with chemokine ligands expressed in lymph nodes (CXCL12) and lung (CCL21), thus guiding metastasis<sup>3,4</sup>.

Tumour-secreted factors can also increase metastasis by inducing vascular leakiness<sup>5</sup>, promoting the recruitment of pro-angiogenic immune cells<sup>6</sup>, and influencing organotropism<sup>7</sup>. Furthermore, the ability of breast cancer to form osteolytic lesions depends on osteoclast-stimulating growth factors (for example, PTHRP and GM-CSF) released into the bone microenvironment<sup>4,8</sup>. Therefore, our previous observation that metastatic melanoma-derived factors dictate organotropism is not surprising<sup>9</sup>. We found that medium conditioned by highly metastatic murine B16-F10 melanoma cells was sufficient to expand the metastatic repertoire of Lewis lung carcinoma cells that would typically metastasize to the lung<sup>9</sup>. We also showed that pre-metastatic niche formation requires S100 protein and fibronectin upregulation by lung resident cells, and the recruitment of bone-marrow-derived

<sup>1</sup>Children's Cancer and Blood Foundation Laboratories, Departments of Pediatrics, and Cell and Developmental Biology, Drukier Institute for Children's Health, Meyer Cancer Center, Weill Cornell Medicine, New York, New York 10021, USA. <sup>2</sup>Department of Plant Pathology and Microbiology and Center for Biotechnology, National Taiwan University, Taipei 10617, Taiwan.

<sup>3</sup>Graduate Program in Areas of Basic and Applied Biology, Abel Salazar Biomedical Sciences Institute, University of Porto, 4099-003 Porto, Portugal. <sup>4</sup>Department of Obstetrics and Gynecology, Faculty of Medicine, University of Tokyo, Tokyo 113-8655, Japan. <sup>5</sup>Proteomics Resource Center, The Rockefeller University, New York, New York 10065, USA. <sup>6</sup>Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA. <sup>7</sup>Department of Oncology and Pathology, Karolinska Institutet, 17176 Stockholm, Sweden. <sup>8</sup>Electron Microscopy Resource Center (EMRC), Rockefeller University, New York, New York 10065, USA. <sup>9</sup>Breast Service, Department of Surgery, Memorial Sloan Kettering Cancer Center, New York, New York, 10065, USA.

<sup>10</sup>Department of Surgery, County Council of Östergötland, and Department of Clinical and Experimental Medicine, Faculty of Health Sciences, Linköping University, 58185 Linköping, Sweden.

<sup>11</sup>Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. <sup>12</sup>Genomics Resources Core Facility, Weill Cornell Medicine, New York, New York 10021, USA.

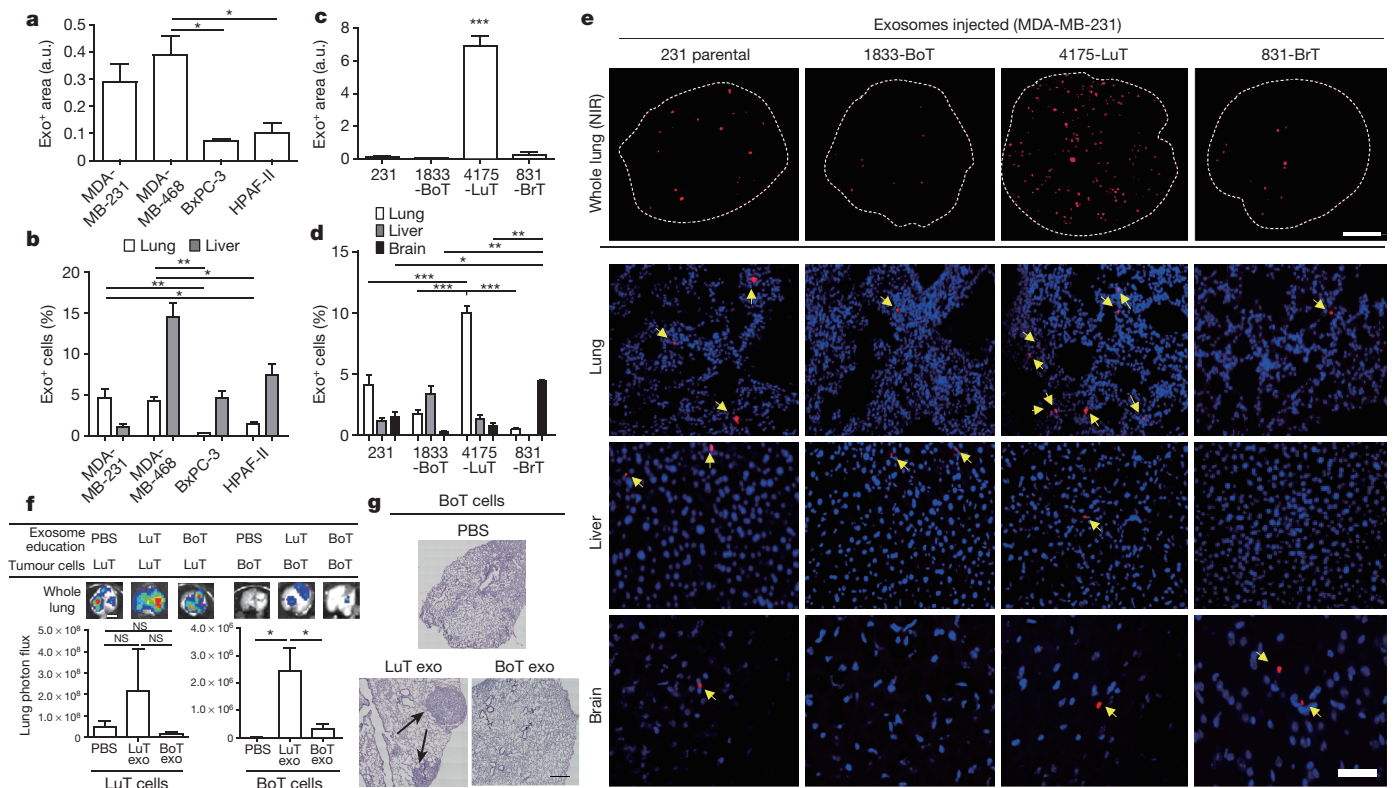
<sup>13</sup>Department of Surgery, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA. <sup>14</sup>Department of Pediatrics, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA. <sup>15</sup>Division of Pediatric Oncology, Alberta Children's Hospital, Calgary, Alberta T3B 6A8, Canada. <sup>16</sup>Division of Hematology/Oncology, Columbia University School of Medicine, New York, New York 10032, USA. <sup>17</sup>Orthopaedic Service, Department of Surgery, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA. <sup>18</sup>Department of Hepato-Pancreato-Biliary Surgery, Oslo University Hospital, Nydalen, Oslo 0424, Norway. <sup>19</sup>Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital, Nydalen, Oslo 0424, Norway.

<sup>20</sup>Eppley Institute for Research in Cancer and Allied Diseases, University of Nebraska Medical Center, Omaha, Nebraska 68198, USA. <sup>21</sup>Department of Biochemistry and Molecular Biology, University of Nebraska Medical Center, Omaha, Nebraska 68198, USA. <sup>22</sup>Gastric and Mixed Tumor Service, Department of Surgery, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA. <sup>23</sup>Department of Tumor Biology, Norwegian Radium Hospital, Oslo University Hospital, Nydalen, Oslo 0424, Norway. <sup>24</sup>Institute for Clinical Medicine, Faculty of Medicine, University of Oslo, Blindern, Oslo 0318, Norway. <sup>25</sup>Department of Gynecology, University Medical Center, Martinistrasse 52, 20246 Hamburg, Germany. <sup>26</sup>Department of Tumor Biology, University Medical Center Hamburg-Eppendorf, Martinistrasse 52, 20246 Hamburg, Germany. <sup>27</sup>Department of Radiation Oncology, Abramson Family Cancer Research Institute, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. <sup>28</sup>Department of Biochemistry and Biophysics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA.

<sup>29</sup>Department of Molecular Biology, Princeton University, Princeton, New Jersey 08544, USA. <sup>30</sup>Rutgers Cancer Institute of New Jersey, New Brunswick, New Jersey 08903, USA. <sup>31</sup>Breast Medicine Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA. <sup>32</sup>Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA.

<sup>33</sup>Microenvironment and Metastasis Laboratory, Department of Molecular Oncology, Spanish National Cancer Research Center (CNIO), Madrid 28029, Spain. <sup>34</sup>Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA. <sup>35</sup>Department of Medicine, Weill Cornell Medicine, New York, New York 10021, USA.

\*These authors contributed equally to this work.



**Figure 1 | Cancer-cell-derived exosomes localize to and dictate future metastatic organs.** **a**, Biodistribution of human cancer-cell-line-derived exosomes in the lung and liver of naive mice. Quantification of exosome-positive (Exo<sup>+</sup>) areas by NIR imaging of whole lung, in arbitrary units (a.u.) ( $n = 3$  per group). **b**, Immunofluorescence quantification of exosome-positive cells ( $n = 3$ , three independent experiments). **c**, MDA-MB-231- (parental), 1833-BoT-, 4175-LuT- and 831-BrT-derived exosome biodistribution. Quantification of exosome-positive areas by NIR imaging of whole lung ( $n = 3$  for all, except 831-BrT, in which  $n = 4$ ). **d**, Immunofluorescence quantification of exosome-positive cells ( $n = 5$  animals pooled from two independent experiments). **e**, Top, NIR whole-lung imaging of MDA-MB-231 sublines. BoT, bone-tropic; BrT, brain-tropic; LuT, lung-tropic. Bottom, myeloid cells in response to tumour-secreted factors<sup>9</sup>. These events establish a favourable microenvironment that promotes the growth of disseminated tumour cells upon their arrival<sup>9–11</sup>.

myeloid cells in response to tumour-secreted factors<sup>9</sup>. These events establish a favourable microenvironment that promotes the growth of disseminated tumour cells upon their arrival<sup>9–11</sup>.

Recently, we demonstrated that exosomes are one of the tumour-derived factors inducing vascular leakiness, inflammation and bone marrow progenitor cell recruitment during pre-metastatic niche formation and metastasis<sup>11</sup>. Exosomes are small membrane vesicles (30–100 nm) containing functional biomolecules (that is, proteins, lipids, RNA and DNA) that can be horizontally transferred to recipient cells<sup>12–19</sup>. We showed that an ‘exosomal protein signature’ could identify melanoma patients at risk for metastasis to nonspecific distant sites<sup>11</sup>. Moreover, in the context of pancreatic cancer exosomes, we defined the sequential steps involved in liver pre-metastatic niche induction<sup>20</sup>.

Taken together, these findings led us to investigate whether molecules present on tumour-derived exosomes are ‘addressing’ them to specific organs. To test this idea, we profiled the exosomal proteome of several tumour models (osteosarcoma, rhabdomyosarcoma, Wilms tumour, skin and uveal melanoma, breast, colorectal, pancreatic and gastric cancers), all of which have a propensity to metastasize to specific sites (that is, brain, lung or liver). We subsequently analysed the biodistribution of tumour-secreted exosomes and found that exosomal integrins (ITGs) direct organ-specific colonization by fusing with target cells in a tissue-specific fashion, thereby initiating pre-metastatic niche formation. Remarkably, we found that tumour-secreted exosomes are sufficient to redirect metastasis of tumour cells that normally lack the capacity to metastasize to a specific organ. Finally, our clinical

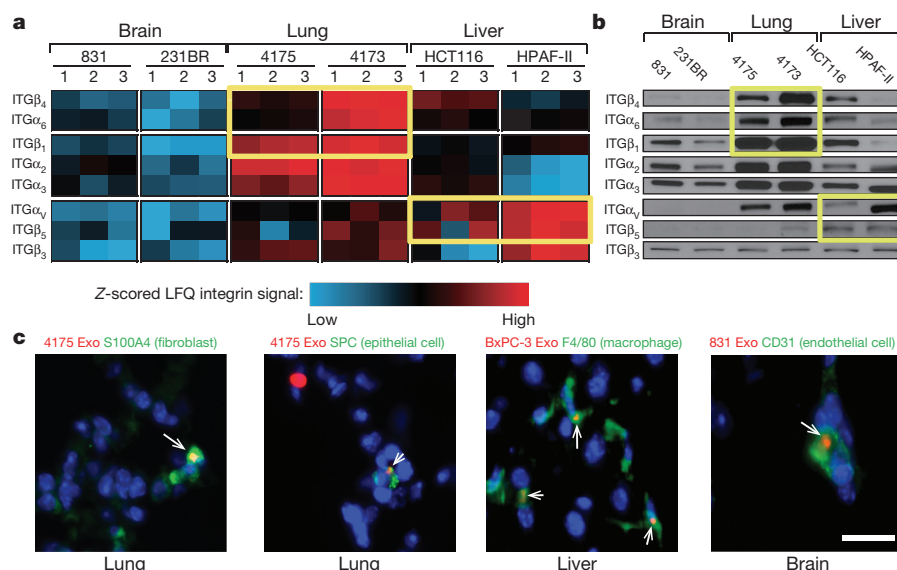
fluorescence microscopy of lung, liver and brain injected with MDA-MB-231 subline-derived exosomes. Arrows indicate exosome foci. All NIR and immunofluorescence images are representative of five random fields. **f**, Redirection of metastasis by education with organotropic exosomes. 4175-LuT or 1833-BoT cell metastasis in the lung after treatment with PBS, 4175-LuT or 1833-BoT exosomes. Top, quantitative bioluminescence of metastatic lesions. Bottom, graphs show quantification of luciferase activity ( $n = 5$  for all, except for LuT exo/LuT cells, in which  $n = 4$ ; data representative of two independent experiments). **g**, Lung haematoxylin/eosin staining for **f**. Arrows indicate lung metastasis. Scale bars, 5 mm (**e**, top, **f**), 50  $\mu$ m (**e**, bottom) and 500  $\mu$ m (**g**). Data are mean  $\pm$  s.e.m. NS, not significant; \* $P < 0.05$ ; \*\* $P < 0.01$ ; \*\*\* $P < 0.001$  by one-way analysis of variance (ANOVA).

data indicate that integrin expression profiles of circulating plasma exosomes isolated from cancer patients could be used as prognostic factors to predict sites of future metastasis. Our findings pave the way for the development of diagnostic tests to predict organ-specific metastasis and therapies to halt metastatic spread.

### Future metastatic sites uptake exosomes

To examine whether tumour exosomes colonize specific organ sites, we isolated exosomes from organotropic human breast and pancreatic cancer cell lines that metastasize primarily to the lung (MDA-MB-231), liver (BxPC-3 and HPAF-II), or both (MDA-MB-468). We then retro-orbitally injected 10  $\mu$ g of near infrared (NIR) or red fluorescently labelled exosomes into nude mice and, 24 h after injection, quantified exosome biodistribution and uptake in distant organs by NIR whole-lung imaging and confocal microscopy (Fig. 1a and Extended Data Fig. 1a). We observed a more than threefold increase in the uptake of MDA-MB-231 and/or 468- versus BxPC-3- and HPAF-II-derived exosomes in the lung (Fig. 1a, b). By contrast, liver uptake of BxPC-3 and HPAF-II exosomes was four times more efficient than that of MDA-MB-231 exosomes (Fig. 1a, b). Moreover, mouse E0771 breast cancer exosomes were four-to-fivefold more efficiently uptaken in lung, whereas mouse Pan02 pancreatic cancer exosomes were four times more efficiently uptaken in liver (Extended Data Fig. 1b). Therefore, the organ specificity of exosome biodistribution matched the organotropic distribution of the cell line of origin in both immune-compromised and immune-competent models.





**Figure 2 | Organ-specific tumour exosomes interact with resident cells.** **a**, Heat map of integrin signals from quantitative mass spectrometry analysis, based on Z-scored label-free quantification (LFQ) values (technical triplicates). **b**, Western blot analysis of integrins from organotropic cell-line-derived exosomes, representative of three independent experiments. For western blot source data, see Supplementary Fig. 1a–h. **c**, Analysis by immunofluorescence of exosome distribution (red) and different resident

cell types (green). Left to right: lung co-staining with 4175-LuT exosomes and S100A4 (fibroblasts) or SPC (epithelial cells), liver co-staining with F4/80 (macrophages) and BxPC-3-LiT exosomes, and brain co-staining with CD31 (endothelial cells) and 831-BrT exosomes. Scale bar, 30  $\mu$ m. Immunofluorescence images are representative of five exosome-positive cells each, from  $n = 5$  mice.

These observations suggested that exosomes could promote organ-specific metastasis. We tested whether exosomes from the MDA-MB-231 sub-lines that colonize lung, bone or brain (4175-LuT, 1833-BoT or 831-BrT, respectively)<sup>21–24</sup> would also exhibit organ tropism. Although exosomes from the MDA-MB-231 variants were similar in size and morphology (Extended Data Fig. 1c), their biodistribution varied 24 h after injection: lung-tropic 4175-LuT exosomes preferentially localized to the lung with a more than fourfold increase in exosome-positive cells compared to 1833-BoT and 831-BrT exosomes (Fig. 1c–e and Extended Data Fig. 1d), whereas 831-BrT exosomes efficiently localized to the brain with a more than fourfold increase compared to 1833-BoT and 4175-LuT exosomes (Fig. 1c–e). Liver and bone showed no significant differences in lung-, brain- or bone-tropic MDA-MB-231-derived exosome distribution, with the exception of 831-BrT exosomes that were uptaken less efficiently by bone marrow cells than exosomes isolated from other MDA-MB-231 sub-lines (Fig. 1d, e and Extended Data Fig. 1e). Taken together, our data suggest that exosomes from different cancer models recapitulate the organ specificity of their cell of origin.

### Exosomes redirect metastatic distribution

We proposed that, in addition to cell-intrinsic genetic determinants of organotropism<sup>23,24</sup>, tumour exosomes could also facilitate organ-specific metastatic behaviour by preparing pre-metastatic niches.

To gain insight into tumour exosome uptake at future metastatic sites, we intravenously injected 4175-LuT exosomes labelled with FM1-43 dye into naive animals, then used electron microscopy to distinguish endogenous from exogenous exosomes in lung sections. We detected tumour FM1-43-labelled exosomes in pre-metastatic cells (Extended Data Fig. 2a; red arrows, exogenous tumour-derived exosomes; black arrows, endogenous stromal exosomes). Moreover, NIR whole-mount lung imaging revealed that NIR-labelled 4175-LuT exosomes accumulated in the lungs of naive animals after three consecutive daily exosome injections (Extended Data Fig. 2b).

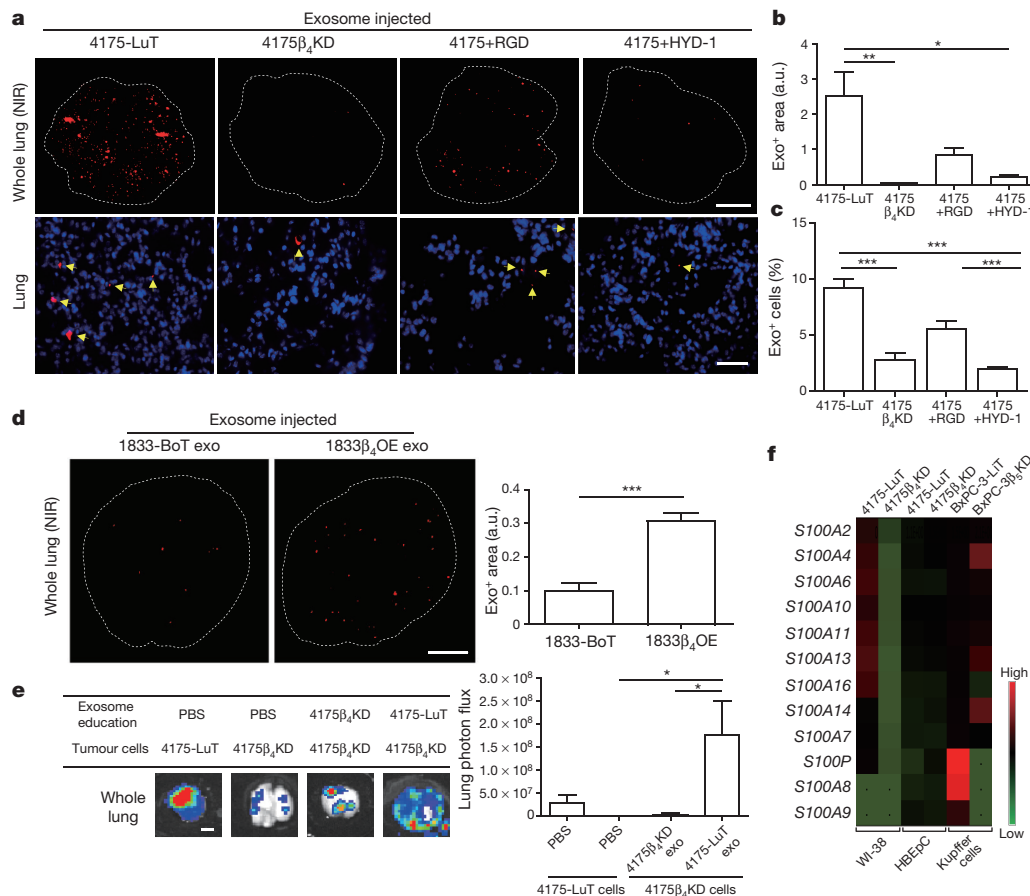
To condition or ‘educate’ cells in specific organs, we retro-orbitally injected 4175-LuT or 1833-BoT exosomes into mice every other day for three weeks<sup>11</sup>. To test exosome education of target organs functionally, luciferase-expressing 4175-LuT or 1833-BoT cells were injected

into exosome-educated mice via the tail vein (Fig. 1f, g and Extended Data Fig. 2c) or intracardially (Extended Data Fig. 2d). Lung-tropic 4175 exosomes marginally increased the lung-metastatic capacity of 4175-LuT tumours. Remarkably, education with 4175-LuT-derived exosomes, but not with 1833-BoT exosomes or PBS, yielded a significant (sevenfold with intravenous and ten-thousand-fold with intracardiac injections) increase in lung metastatic capacity of 1833-BoT cells (Fig. 1f, g and Extended Data Fig. 2d). Conversely, 1833-BoT-derived exosomes did not affect 4175-LuT cell metastasis to the lung (Fig. 1f and Extended Data Fig. 2c). These data suggest that organotropic tumour exosomes prepare pre-metastatic niches potent enough to facilitate metastasis even for tumour cells poorly capable of colonizing these sites.

### Exosomal ITGs determine organotropism

We then postulated that exosomal adhesion molecules could regulate local microenvironments within future metastatic organs. Quantitative mass spectrometry of brain-, lung- and liver-tropic metastatic exosomes identified six integrins among the top 40 most abundant adhesion molecules, making integrins the most highly represented protein family in this analysis. These data indicate a correlation between exosomal integrins and metastatic tropism (Extended Data Fig. 3a).

Interestingly, we found that integrin expression profiles correlated with tissue organotropism. Both quantitative mass spectrometry (Fig. 2a) and western blot analysis (Fig. 2b and Extended Data Fig. 3b) revealed that integrin  $\alpha_6$  (ITG $\alpha_6$ ), and its partners ITG $\beta_4$  and ITG $\beta_1$  (ref. 25), were present abundantly in lung-tropic exosomes. By contrast, ITG $\beta_5$ , which associates only with ITG $\alpha_v$  (ref. 25), was detected primarily in liver-tropic exosomes (Fig. 2a, b). We confirmed these findings by exosome proteomics for 28 organ-specific metastatic cell lines (Extended Data Tables 1 and 2). Qualitative mass spectrometry revealed that ITG $\alpha_6$  was present in lung-tropic exosomes, whereas ITG $\beta_5$  was found in liver-tropic exosomes (Extended Data Tables 1 and 2), consistent with our quantitative proteomics data. Exosomes from 4173, 4175 and 4180 lung-tropic MDA-MB-231 variants expressed ITG $\alpha_6\beta_4$  (Extended Data Table 1). Meanwhile, ITG $\beta_3$  was present in exosomes isolated from brain-tropic cells (Extended Data Table 1). Notably, unlike non-cancerous lung fibroblast WI-38



**Figure 3 | Exosomal ITG $\beta_4$  expression functionally contributes to 4175-LuT exosome localization and mediates lung metastasis.** **a**, Top, NIR whole-lung imaging of 4175-LuT- or 4175 $\beta_4$ KD-derived exosomes, or 4175-LuT-derived exosomes pre-incubated with RGD or HYD-1 peptides. Bottom, fluorescence microscopy. Arrowheads indicate exosome foci. **b**, Quantification of exosome-positive areas from the whole-lung images in **a** (top) ( $n = 4$ , except 4175, in which  $n = 6$ ). **c**, Immunofluorescence quantification of exosome-positive cells from **a** (bottom) ( $n = 6$  pooled from two independent experiments). **d**, Left, NIR whole-lung imaging of 1833-BoT ( $n = 5$ ) or 1833-BoT overexpressing ITG $\beta_4$  (1833 $\beta_4$ OE) ( $n = 4$ ) exosomes. Right, quantification of the exosome-positive areas. **e**, Experimental lung metastasis of 4175 $\beta_4$ KD cells after education with wild-

or epithelial MCF10A exosomes, metastatic cell exosomes contained ITG $\alpha_2\beta_1$ , suggesting that this integrin could serve as a biomarker for metastasis (Extended Data Table 1). Importantly, exosomal integrin expression does not necessarily reflect cellular integrin expression, consistent with selective packaging of integrins in exosomes (Extended Data Fig. 3c). Taken together, our data suggest that exosomal integrin expression patterns underlie organotropism to the lung, liver and brain.

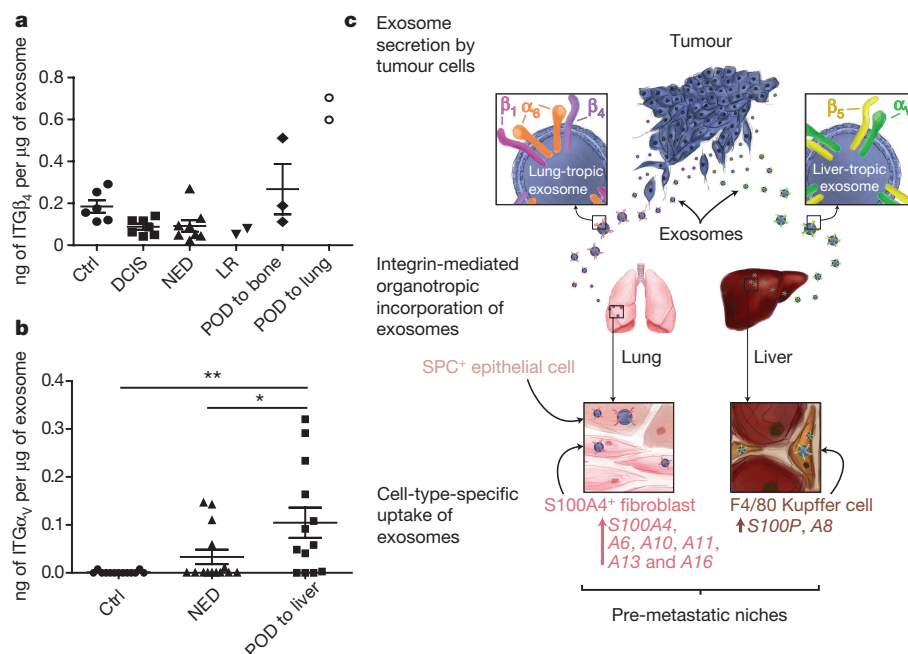
### Distinct cells uptake tropic exosomes

To identify the cells uptaking tumour exosomes in each organ, we intravenously injected red fluorescently labelled exosomes from 4175-LuT, 1833-BoT, BxPC-3-liver-tropic (BxPC-3-LiT) or 831-BrT cells into mice (Fig. 2c and Extended Data Fig. 4a, b). Both 1833-BoT and 4175-LuT exosomes promoted vascular leakiness 24 h after injection, before exosome uptake by specific lung cells (Extended Data Fig. 4b). These observations fit with our previous studies using melanoma exosomes<sup>11</sup>, suggesting that exosomes first permeabilize vessels, allowing for exosome diffusion before uptake by parenchymal cells. Unexpectedly, we found that the specific cell type responsible for exosome uptake varied depending on the metastatic organ. Lung-tropic 4175 exosomes mainly co-localized

type or 4175 $\beta_4$ KD exosomes. Bioluminescence imaging of lung metastasis and quantification of luciferase activity ( $n = 6$ , data representative of two independent experiments). **f**, Heat map of S100 gene expression fold change by quantitative reverse transcription PCR (qRT-PCR) in 4175-LuT or 4175 $\beta_4$ KD exosome-conditioned lung fibroblast (WI-38) or epithelial (HBEPC) cells, and liver-tropic BxPC-3 or BxPC-3 $\beta_5$ KD exosome-conditioned Kupffer cells. Red represents high and green represents low expression ( $n = 3$  in two independent experiments). Scale bars, 5 mm (**a**, top, **d**, **e**) and 50  $\mu$ m (**b**, bottom). Data are mean  $\pm$  s.e.m. \* $P < 0.05$ ; \*\* $P < 0.01$ ; \*\*\* $P < 0.001$  by one-way ANOVA (**b**, **c**, **e**); \*\*\* $P < 0.001$  by two-tailed Student's *t*-test (**d**).

with S100A4-positive fibroblasts and surfactant protein C (SPC)-positive epithelial cells (40% and 30% of exosome-positive cells, respectively) in the lung (Fig. 2c and Extended Data Figs 4a and 5a, top). By contrast, pancreatic cancer exosomes derived from BxPC-3-LiT cells fused with Kupffer cells<sup>20</sup> (90% of exosome-positive cells; Fig. 2c and Extended Data Figs 4a and 5a, middle). Finally, 831-BrT exosomes interacted mainly with CD31-positive brain endothelial cells (98% of exosome-positive cells; Fig. 2c and Extended Data Figs 4a and 5a, bottom). Collectively, these data demonstrate that specific tissue-resident stromal cells differentially uptake tumour exosomes in metastatic target organs.

We proposed that the unique exosomal integrins may interact with cell-associated extracellular matrix (ECM), mediating exosome uptake in specific target organs. We found that 4175-LuT exosomes expressing ITG $\alpha_6\beta_4$  and ITG $\alpha_6\beta_1$  co-localized with S100A4-positive cells in laminin-rich lung microenvironments (Extended Data Fig. 5b, top). Meanwhile, ITG $\alpha_v\beta_5$ -expressing pancreatic BxPC-3-LiT exosomes co-localized with F4/80<sup>+</sup> macrophages in fibronectin-rich liver microenvironments (Extended Data Fig. 5b, bottom). Therefore, specific exosomal integrins may selectively adhere to ECM-enriched cellular areas in the lung and liver.



**Figure 4 | Exosomal integrin expression as a potential predictor of patient organ-specific metastasis.** **a**, Exosomal ITGβ<sub>4</sub> levels in breast cancer patients who were metastasis-free at the time of blood draw. Amount of ITGβ<sub>4</sub> per microgram of exosome in healthy control (Ctrl) subjects ( $n=6$ ); patients with ductal carcinoma *in situ* (DCIS) ( $n=7$ ), invasive breast cancer without relapse within three years (NED, no evidence of disease) ( $n=8$ ), locoregional recurrence (LR) within three years ( $n=2$ ), bone metastasis within three years ( $n=3$ ), or lung metastasis within three years ( $n=2$ ). POD, progression of disease. **b**, Exosomal ITGα<sub>v</sub> in pancreatic

cancer patients who were metastasis-free at the time of blood draw. Amount of ITGα<sub>v</sub> per microgram of exosome in healthy control subjects ( $n=13$ ); patients with pancreatic cancer without relapse within three years ( $n=14$ ), or liver metastasis within three years ( $n=13$ ). **c**, Model of exosome-mediated organotropic tumour dissemination. Tumour-derived exosomes are taken up by organ-specific resident cells in future metastatic organs based on integrin expression. Data are mean  $\pm$  s.e.m. \* $P < 0.05$ ; \*\* $P < 0.01$  by one-way ANOVA.

### Exosomal tropism requires ITGβ<sub>4</sub> and ITGβ<sub>5</sub>

We next asked whether manipulating the integrin cargo packaged into exosomes could impact metastatic organotropism. To test the requirement for exosomal ITGβ<sub>4</sub> in lung tropism, we knocked down ITGβ<sub>4</sub> expression in 4175-LuT cells using short hairpin RNAs (shRNAs) (4175β<sub>4</sub>KD; Extended Data Fig. 6a). We found a more than threefold reduction in labelled ITGβ<sub>4</sub>KD exosomes in the lung compared with labelled control exosomes 24 h after injection (Fig. 3a–c). To test the requirement for exosomal ITGβ<sub>4</sub> binding to laminin for lung tropism, we blocked integrin binding using RGD and HYD-1 peptides. Pre-incubation of 4175-LuT exosomes with HYD-1, which blocks laminin receptors<sup>26</sup>, markedly reduced exosome uptake in the lung, whereas RGD, which blocks fibronectin receptors but not ITGβ<sub>4</sub> (ref. 27), did not significantly alter exosome uptake in the lung (Fig. 3a–c). Pre-treatment of 4175-LuT exosomes with HYD-1 peptide also prevented their uptake by WI-38 lung fibroblasts *in vitro* (Extended Data Fig. 6b). Conversely, ITGβ<sub>4</sub> overexpression in 1833-BoT exosomes was sufficient to increase exosome uptake in lung (Extended Data Fig. 6c and Fig. 3d). These data demonstrate that integrins are responsible for organ-specific uptake of exosomes, and that ITGβ<sub>4</sub> promotes tumour exosome adhesion within the lung.

We next investigated whether ITGβ<sub>4</sub>KD exosomes could modulate the metastatic organotropism of 4175-LuT models. Knockdown of ITGβ<sub>4</sub> was sufficient to reduce the lung metastatic capacity of 4175-LuT cells (Fig. 3e). Education with 4175-LuT exosomes, but not ITGβ<sub>4</sub>KD exosomes, rescued the metastatic ability of ITGβ<sub>4</sub>KD cells, yielding metastasis similar to 4175-LuT cells (Fig. 3e and Extended Data Fig. 6d). Therefore, ITGβ<sub>4</sub>-expressing exosomes can confer lung-metastatic behaviour to cells with limited capacity to colonize the lung. Similarly, ITGβ<sub>5</sub> knockdown in BxPC-3-LiT exosomes decreased liver uptake by sevenfold compared with control BxPC-3-LiT exosomes (Extended Data Fig. 6e, f). Moreover,

pre-incubation with RGD peptide or anti-ITGα<sub>v</sub>β<sub>5</sub> antibody, but not HYD-1 peptide, significantly reduced BxPC-3-LiT and Pan02-LiT exosome adhesion to the liver (Extended Data Figs 6g and 7a, respectively). Importantly, RGD peptides also inhibited the education effect of Pan02-LiT exosomes, subsequently blocking pre-metastatic niche formation and liver metastasis (Extended Data Fig. 7b). Our data support the hypothesis that local microenvironmental changes induced by specific exosomal cargo (that is, ITGβ<sub>4</sub> or ITGβ<sub>5</sub>) can dictate metastatic organotropism.

### Exosomal ITGs activate S100 genes

To identify the downstream effects of exosomal interaction with target cells, Kupffer cells were educated with either BxPC-3-LiT or BxPC-3-LiT ITGβ<sub>5</sub>KD exosomes every other day for two weeks. Unbiased analysis of gene expression by RNA sequencing in Kupffer cells identified 906 genes upregulated more than twofold after treatment with BxPC-3-LiT exosomes compared to BxPC-3-LiT ITGβ<sub>5</sub>KD exosomes. Cell migration genes were the most prominently upregulated (twofold for 221 genes; fourfold for 42 genes). Of these, *S100A8* and *S100P* were upregulated more than fourfold (Fig. 3f; GEO accession GSE68919). Since pro-inflammatory *S100* gene expression correlates with metastasis<sup>28,29</sup>, we analysed *S100* genes in tumour exosome-educated lung WI-38 fibroblasts and in human bronchial epithelial cells (HBEPs). Several *S100* genes (*S100A4*, -A6, -A10, -A11, -A13 and -A16) were upregulated more than fivefold after WI-38 fibroblast treatment with 4175-LuT exosomes compared with 4175-LuT ITGβ<sub>4</sub>KD exosomes (Fig. 3f). Notably, *S100* genes remained unchanged in HBEPs treated with 4175-LuT exosomes (Fig. 3f). Moreover, exosome-treated lung fibroblasts proliferated and migrated more than controls (Extended Data Fig. 7c, d), which correlated with a higher frequency of S100A4<sup>+</sup> cells in the lungs after three weeks of education with 4175-LuT, but not 4175β<sub>4</sub>KD, exosomes (Extended Data Fig. 7e). We then surveyed, by in-cell western blot analysis, ITGβ<sub>4</sub>



signalling proteins<sup>30–34</sup> in WI-38 fibroblasts treated with 4175-LuT or 4175 $\beta_4$ KD exosomes. Notably, only Src or phosphorylated Src (pSrc) levels increased in an exosomal ITG $\beta_4$ -dependent manner (Extended Data Fig. 7f), consistent with the known roles of ITG $\alpha_6\beta_4$  in Src activation and S100A4 expression<sup>30</sup>. Therefore, in addition to their adhesive properties, exosomal integrins can activate Src and upregulate pro-migratory and pro-inflammatory S100 molecules in specific resident cells within distant tissue microenvironments, influencing the expression of genes implicated in facilitating tumour metastasis.

### Exosomal ITGs as organotropism biomarkers

Next we investigated whether exosomal integrin content could predict tumour progression. ITG $\beta_4$  levels were increased in the plasma of mice six weeks after orthotopic 4175-LuT cell injection into the mammary fat pad, but were significantly reduced after successful tumour resection (Extended Data Fig. 8a). Furthermore, we performed ELISA assays for plasma-derived exosomal integrins in patients with lung (ITG $\beta_4$ ) or liver (ITG $\alpha_v$ , the binding partner of ITG $\beta_5$ ) metastasis. We found increased ITG $\beta_4$  levels in exosomes from patients with lung metastasis (regardless of tumour type) compared with patients with no metastasis or liver metastasis (Extended Data Fig. 8b). Exosomes isolated before metastasis from patients with breast cancer that progressed to develop lung metastasis (POD) expressed the highest levels of exosomal ITG $\beta_4$  (Fig. 4a). ITG $\alpha_v$  was significantly increased in exosomes isolated from cancer patients with liver metastasis compared with patients with no metastasis or lung metastasis (Extended Data Fig. 8c). Finally, exosomal ITG $\alpha_v$  levels at diagnosis were higher in patients with pancreatic cancer who developed liver metastasis than in those without liver metastasis within three years of diagnosis or in control subjects (Fig. 4b). Taken together, our data indicate that the specific exosomal integrins in breast and pancreatic cancer patient plasma correlate with and predict likely sites of metastasis.

### Discussion

Since Stephen Paget's hypothesis first emerged, many studies have focused on identifying cell-intrinsic determinants of organ-specific metastasis<sup>3,7,23,24,35</sup>. We now show that tumour-derived exosomes prepare a favourable microenvironment at future metastatic sites and mediate non-random patterns of metastasis. We identify determinants of exosome-mediated organ-specific conditioning that allow the redirection of metastasis. Previously, adhesion and ECM molecules, such as integrins, tenascin and periostin, were shown to promote metastasis of disseminating cancer cells<sup>36–39</sup>. We define a specific repertoire of integrins expressed on tumour-derived exosomes, distinct from tumour cells, which dictates exosome adhesion to specific cell types and ECM molecules in particular organs. Notably, exosomes expressing ITG $\alpha_v\beta_5$  specifically bind to Kupffer cells, mediating liver tropism, whereas exosomal ITG $\alpha_6\beta_4$  and ITG $\alpha_6\beta_1$  bind lung-resident fibroblasts and epithelial cells, governing lung tropism (Fig. 4c).

Interestingly, bone-tropic exosomes expressed a limited integrin repertoire, but were capable of inducing vascular leakiness in the lung despite lack of uptake in the lung parenchyma. These results suggest that whereas induction of vascular leakiness may be the first exosome-mediated step during the metastatic cascade, it is insufficient to promote metastasis. Thus, integrin-independent mechanisms may mediate vascular leakiness and exosome involvement in bone metastasis.

Cell-type-specific exosome integrin uptake promoted pro-migratory and pro-inflammatory S100 gene upregulation (S100A4, -A6, -A10, -A11, -A13 and -A16 in lung fibroblasts; S100P and -A8 in Kupffer cells). Notably, tumour exosomes failed to elicit S100 upregulation in lung epithelial cells, highlighting the cell-type specificity of exosomal education. Since S100A4 regulates lung metastasis<sup>40</sup> and is controlled by ITG $\alpha_6\beta_4$  (ref. 41), we conclude that exosomal ITG $\alpha_6\beta_4$  activates the Src–S100A4 axis in lung fibroblasts during pre-metastatic niche

formation. Therefore, we propose that exosomal integrins not only promote adhesion, but also trigger signalling pathways and inflammatory responses in target cells resulting in the education of that organ and rendering it permissive for the growth of metastatic cells.

We provide the proof-of-principle that integrin-blocking decoy peptides successfully ablate tumour exosome adhesion in an integrin-specific and organ-specific manner. Thus, it is no longer surprising that targeting ITG $\alpha_v$  in breast cancer cells prevented metastasis to other organs but not to the lung<sup>42–44</sup>. However, strategies targeting exosomal integrins may effectively block organ-specific metastasis. Collectively, our data suggest that exosomal integrins and exosome-inducible S100 molecules in target cells represent candidates for anti-metastatic combination therapies.

Overall, our findings suggest that circulating tumour-derived exosomes may be useful not only to predict metastatic propensity<sup>7</sup>, but also to determine organ sites of future metastasis. We believe exosomes perform distinct roles during each of the sequential steps (that is, vascular leakiness, stromal cell education at organotropic sites, bone-marrow-derived cell education and recruitment) necessary to complete pre-metastatic niche evolution<sup>11,20,45</sup>.

Future studies will focus on identifying exosomal integrins and proteins that could dictate metastasis to other organs, as well as further exploring the potential of exosomal ITG $\alpha_2\beta_1$  as a marker and driver of all cancer metastasis. Our findings demonstrate an important role for exosomes in dictating organ-specific metastasis, thus providing a basis for deciphering the mystery of organotropism.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 30 October 2014; accepted 29 September 2015.

Published online 28 October 2015.

- Paget, S. The distribution of secondary growths in cancer of the breast. 1889. *Cancer Metastasis Rev.* **8**, 98–101 (1989).
- Hart, I. R. & Fidler, I. J. Role of organ selectivity in the determination of metastatic patterns of B16 melanoma. *Cancer Res.* **40**, 2281–2287 (1980).
- Müller, A. *et al.* Involvement of chemokine receptors in breast cancer metastasis. *Nature* **410**, 50–56 (2001).
- Weilbaecher, K. N., Guise, T. A. & McCauley, L. K. Cancer to bone: a fatal attraction. *Nature Rev. Cancer* **11**, 411–425 (2011).
- Zhou, W. *et al.* Cancer-secreted miR-105 destroys vascular endothelial barriers to promote metastasis. *Cancer Cell* **25**, 501–515 (2014).
- Chang, Q. *et al.* The IL-6/JAK/Stat3 feed-forward loop drives tumorigenesis and metastasis. *Neoplasia* **15**, 848–862 (2013).
- Lu, X. & Kang, Y. Organotropism of breast cancer metastasis. *J. Mammary Gland Biol. Neoplasia* **12**, 153–162 (2007).
- Cox, T. R. *et al.* The hypoxic cancer secretome induces pre-metastatic bone lesions through lysyl oxidase. *Nature* **522**, 106–110 (2015).
- Kaplan, R. N. *et al.* VEGFR1-positive haematopoietic bone marrow progenitors initiate the pre-metastatic niche. *Nature* **438**, 820–827 (2005).
- Hiratsuka, S. *et al.* MMP9 induction by vascular endothelial growth factor receptor-1 is involved in lung-specific metastasis. *Cancer Cell* **2**, 289–300 (2002).
- Peinado, H. *et al.* Melanoma exosomes educate bone marrow progenitor cells toward a pro-metastatic phenotype through MET. *Nature Med.* **18**, 883–891 (2012).
- Balaj, L. *et al.* Tumour microvesicles contain retrotransposon elements and amplified oncogene sequences. *Nature Commun.* **2**, 180 (2011).
- Skog, J. *et al.* Glioblastoma microvesicles transport RNA and proteins that promote tumour growth and provide diagnostic biomarkers. *Nature Cell Biol.* **10**, 1470–1476 (2008).
- Théry, C., Ostrowski, M. & Segura, E. Membrane vesicles as conveyors of immune responses. *Nature Rev. Immunol.* **9**, 581–593 (2009).
- Raposo, G. & Stoorvogel, W. Extracellular vesicles: exosomes, microvesicles, and friends. *J. Cell Biol.* **200**, 373–383 (2013).
- Peinado, H., Lavotshkin, S. & Lyden, D. The secreted factors responsible for pre-metastatic niche formation: old sayings and new thoughts. *Semin. Cancer Biol.* **21**, 139–146 (2011).
- Choi, D. S., Kim, D. K., Kim, Y. K. & Gho, Y. S. Proteomics, transcriptomics and lipidomics of exosomes and ectosomes. *Proteomics* **13**, 1554–1571 (2013).
- Valadi, H. *et al.* Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells. *Nature Cell Biol.* **9**, 654–659 (2007).

19. Thakur, B. K. *et al.* Double-stranded DNA in exosomes: a novel biomarker in cancer detection. *Cell Res.* **24**, 766–769 (2014).
20. Costa-Silva, B. *et al.* Pancreatic cancer exosomes initiate pre-metastatic niche formation in the liver. *Nature Cell Biol.* **17**, 816–826 (2015).
21. Kang, Y. *et al.* A multigenic program mediating breast cancer metastasis to bone. *Cancer Cell* **3**, 537–549 (2003).
22. Gupta, G. P. *et al.* Identifying site-specific metastasis genes and functions. *Cold Spring Harb. Symp. Quant. Biol.* **70**, 149–158 (2005).
23. Minn, A. J. *et al.* Genes that mediate breast cancer metastasis to lung. *Nature* **436**, 518–524 (2005).
24. Bos, P. D. *et al.* Genes that mediate breast cancer metastasis to the brain. *Nature* **459**, 1005–1009 (2009).
25. Desgrosellier, J. S. & Cheresh, D. A. Integrins in cancer: biological implications and therapeutic opportunities. *Nature Rev. Cancer* **10**, 9–22 (2010).
26. Sroka, T. C., Marik, J., Pennington, M. E., Lam, K. S. & Cress, A. E. The minimum element of a synthetic peptide required to block prostate tumor cell migration. *Cancer Biol. Ther.* **5**, 1556–1562 (2006).
27. Ruoslahti, E. & Pierschbacher, M. D. Arg-Gly-Asp: a versatile cell recognition signal. *Cell* **44**, 517–518 (1986).
28. Grum-Schwensen, B. *et al.* Suppression of tumor development and metastasis formation in mice lacking the *S100A4* (*mts1*) gene. *Cancer Res.* **65**, 3772–3780 (2005).
29. Lukanidin, E. & Sleeman, J. P. Building the niche: the role of the S100 proteins in metastatic growth. *Semin. Cancer Biol.* **22**, 216–225 (2012).
30. Kim, T. H., Kim, H. I., Soung, Y. H., Shaw, L. A. & Chung, J. Integrin ( $\alpha 6 \beta 4$ ) signals through Src to increase expression of S100A4, a metastasis-promoting factor: implications for cancer cell invasion. *Mol. Cancer Res.* **7**, 1605–1612 (2009).
31. Abdel-Ghany, M., Cheng, H. C., Elble, R. C. & Pauli, B. U. Focal adhesion kinase activated by  $\beta_4$  integrin ligation to mCLCA1 mediates early metastatic growth. *J. Biol. Chem.* **277**, 34391–34400 (2002).
32. Mainiero, F. *et al.* p38 MAPK is a critical regulator of the constitutive and the beta4 integrin-regulated expression of IL-6 in human normal thymic epithelial cells. *Eur. J. Immunol.* **33**, 3038–3048 (2003).
33. Weaver, V. M. *et al.*  $\beta 4$  integrin-dependent formation of polarized three-dimensional architecture confers resistance to apoptosis in normal and malignant mammary epithelium. *Cancer Cell* **2**, 205–216 (2002).
34. Nikolopoulos, S. N. *et al.* Targeted deletion of the integrin beta4 signaling domain suppresses laminin-5-dependent nuclear entry of mitogen-activated protein kinases and NF- $\kappa$ B, causing defects in epidermal growth and migration. *Mol. Cell. Biol.* **25**, 6090–6102 (2005).
35. Minn, A. J. *et al.* Distinct organ-specific metastatic potential of individual breast cancer cells and primary tumors. *J. Clin. Invest.* **115**, 44–55 (2005).
36. Oskarsson, T. *et al.* Breast cancer cells produce tenascin C as a metastatic niche component to colonize the lungs. *Nature Med.* **17**, 867–874 (2011).
37. Fukuda, K. *et al.* Periostin is a key niche component for wound metastasis of melanoma. *PLoS ONE* **10**, e0129704 (2015).
38. Radisky, D., Muschler, J. & Bissell, M. J. Order and disorder: the role of extracellular matrix in epithelial cancer. *Cancer Invest.* **20**, 139–153 (2002).
39. Weaver, V. M. *et al.* Reversion of the malignant phenotype of human breast cells in three-dimensional culture and *in vivo* by integrin blocking antibodies. *J. Cell Biol.* **137**, 231–245 (1997).
40. Grum-Schwensen, B. *et al.* Lung metastasis fails in MMTV-PyMT oncomice lacking S100A4 due to a T-cell deficiency in primary tumors. *Cancer Res.* **70**, 936–947 (2010).
41. Chen, M., Sinha, M., Luxon, B. A., Bresnick, A. R. & O'Connor, K. L. Integrin  $\alpha 6 \beta 4$  controls the expression of genes associated with cell motility, invasion, and metastasis, including S100A4/metastasin. *J. Biol. Chem.* **284**, 1484–1494 (2009).
42. Bäuerle, T. *et al.* Cilengitide inhibits progression of experimental breast cancer bone metastases as imaged noninvasively using VCT, MRI and DCE-MRI in a longitudinal *in vivo* study. *Int. J. Cancer* **128**, 2453–2462 (2011).
43. Wu, Y. J. *et al.* Targeting  $\alpha$ V-integrins decreased metastasis and increased survival in a nude rat breast cancer brain metastasis model. *J. Neurooncol.* **110**, 27–36 (2012).
44. Zhao, Y. *et al.* Tumor  $\alpha_6 \beta_3$  integrin is a therapeutic target for breast cancer bone metastases. *Cancer Res.* **67**, 5821–5830 (2007).
45. Tominaga, N. *et al.* Brain metastatic cancer cells release microRNA-181c-containing extracellular vesicles capable of destructing blood-brain barrier. *Nature Commun.* **6**, 6716 (2015).

**Supplementary information** is available in the online version of the paper.

**Acknowledgements** We thank S. Rudchenko at the Hospital for Special Surgery Flow Cytometry Core Facility. We acknowledge the MSK Cancer Center Support Grant/Core Grant (P30 CA008748). Our work is supported by grants from National Cancer Institute (U01-CA169538, D.L. and M.S.B.), National Institutes of Health (R01-CA169416, D.L. and H.P.), United States Department of Defense (W81XWH-13-10249, D.L.), W81XWH-13-1-0425 (D.L., J.B., B.A.G. and Y.K.), Melanoma Research Alliance (H.P.), Sohn Conference Foundation (H.P. and H.Z.), the Children's Cancer and Blood Foundation (H.P. and D.L.), The Manning Foundation (D.L.), The Hartwell Foundation (D.L.), Fundação para a Ciência e a Tecnologia (D.L.), The Nancy C. and Daniel P. Paduano Foundation (H.P. and D.L.), The Feldstein Foundation (H.P.), The Starr Cancer Consortium (H.P. and D.L.), The Mary Kay Foundation (D.L.), Pediatric Oncology Experimental Therapeutic Investigator Consortium (POETIC, D.L. and H.P.), James Paduano Foundation (D.L. and H.P.), Beth Tortolani Foundation (D.L. and J.B.), Malcolm Hewitt Weiner Foundation (D.L.), Theodore A. Rapp Foundation (D.L.), American Hellenic Educational Progressive Association 5th District Cancer Research Foundation (D.L., A.H.), Charles and Marjorie Holloway Foundation (J.B.), Sussman Family Fund (J.B.), Lerner Foundation (J.B.), Breast Cancer Alliance (J.B.), Manhasset Women's Coalition Against Breast Cancer (J.B.), Ministry of Science and Technology Taiwan (101-2918-I-002-016, T.-L.S.), The JSPS Postdoctoral Fellowships for Research Abroad and Susan G. Komen Postdoctoral Fellowship (A.H.).

**Author Contributions** A.H. designed the experimental approach, performed the experimental work, analysed the data, coordinated the project and wrote the manuscript. B.C.-S. designed experiments investigating liver metastasis and performed the experimental work. T.-L.S. performed ECM studies. G.R. analysed brain tropic exosome distribution. A.H. performed western blot analysis. M.T.M. and H.M. performed and analysed exosome mass spectrometry. S.K. prepared overexpression vectors. S.S. and L.B. performed tissue processing and staining. S.C. designed and illustrated Fig. 4c. A.D.G., S.C., V.D.D.-C., Y.A. and C.W. received and processed human samples. N.S. and K.U. performed electron microscopy. A.E.D. performed animal surgeries and contributed to data interpretation and discussion. T.Z. performed RNA sequence analysis. B.A.G. performed initial proteomic analysis. V.K.R., G.K.S. and J.H.H. provided the uveal melanoma cell line. L.P., T.K., M.S.B., V.M., K.K., L.H.W., J.H., E.H.K., K.M., S.K.B., K.P., O.F., M.J., S.K., M.A.H., P.M.G., K.J.L., J.M.W., A.N. and W.R.J. provided and prepared human samples. H.Z., A.J.M. and P.S. read the manuscript and provided feedback. C.M.G., I.M. and H.P. discussed the hypothesis and contributed to data interpretation and wrote the manuscript. Y.K., M.d.S. and M.J.B. contributed to discussing the hypothesis, interpretation of data. J.B. coordinated the project, interpreted data and wrote the manuscript. D.L. conceived the hypothesis, led the project, interpreted the data and wrote the manuscript.

**Author Information** The raw data for quantitative mass spectrometry analysis of lung-tropic (4173 and 4175), liver-tropic (HPAF-II and HCT116) and brain-tropic (831 and 231BR) exosomes (Fig. 2a and Extended Data Fig. 3a) are available at <http://dx.doi.org/10.6084/m9.figshare.1569781>. The raw sequencing data for human Kupffer cells treated *in vitro* with BxPC-3 or BxPC-3 ITG $\beta_5$ KD exosomes have been deposited in the Gene Expression Omnibus (GEO) under accession number GSE68919. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to H.P. ([hpeinado@cniio.es](mailto:hpeinado@cniio.es)), J.B. ([bromberj@mskcc.org](mailto:bromberj@mskcc.org)) or D.L. ([dcl2001@med.cornell.edu](mailto:dcl2001@med.cornell.edu)).

## METHODS

**Cell lines and cell culture.** The cell lines used in this study were provided as follows: human breast cancer MDA-MB-231 organotropic lines 4175, 1833 and 831 by J. Massagué; human breast cancer 4173 and 4180 cells by A. Minn; human breast cancer 231BR cells by P. Steeg; liver metastasis enriched uveal melanoma cells by V. Rajasekhar; human osteosarcoma 143B cells by A. Narendran; human melanoma 131/4-5B2 and 131/8-2L cells by R. Kerbel; human melanoma SB1B cells by C. E. Verschraegen; human rhabdomyosarcoma CT10 and RD cells by R. Gladdy; and human Wilms tumour CCG9911 and CLS1 cells by A. Ketsis. Human breast cancer cell lines MDA-MB-231 and MDA-MB-468, human breast epithelial cells MCF10A, human pancreatic cancer cell lines, gastric cancer cell lines and colorectal cancer cell lines were purchased from American Type Culture Collection (ATCC). Although HT29 is commonly misidentified, we purchased this cell line directly from ATCC and the cell line was certified by this repository, therefore we are confident that it is indeed a colon cancer cell line. The C57BL/6 mouse pancreatic adenocarcinoma Pan02 was purchased from the National Cancer Institute Tumour Repository (DTP/DCTD, Frederick National Laboratory for Cancer Research). For *in vitro* education of human lung fibroblasts WI-38 (ATCC), human bronchial epithelial cells HBEpC (PromoCell), and human Kupffer cells (Life Technologies), cells were maintained in culture for 14 days, with media containing 0, 5 or 10  $\mu\text{g ml}^{-1}$  of exosomes, replenished every other day. Kupffer cells were cultured in RPMI and WI-38 cells were cultured in alpha-MEM, both supplemented with 10% exosome-depleted FBS (Gibco, Thermo Fisher Scientific) and penicillin-streptomycin. HBEpC cells were cultured in airway epithelial cell growth medium (PromoCell). All cells were maintained in a humidified incubator with 5%  $\text{CO}_2$  at 37°C. FBS was depleted of bovine exosomes by ultracentrifugation at 100,000g for 70 min. All cell lines were routinely tested for mycoplasma and were found to be negative.

**Exosome purification, characterization and analyses.** Exosomes were purified by sequential centrifugation as previously described<sup>46</sup>. In brief, cells were removed from 3–4-day cell culture supernatant by centrifugation at 500g for 10 min to remove any cell contamination. To remove any possible apoptotic bodies and large cell debris, the supernatants were then spun at 12,000g for 20 min. Finally, exosomes were collected by spinning at 100,000g for 70 min. Exosomes were washed in 20 ml PBS and pelleted again by ultracentrifugation (Beckman 70Ti rotor). Exosome preparations were verified by electron microscopy. Exosome size and particle number were analysed using the LM10 or DS500 nanoparticle characterization system (NanoSight, Malvern Instruments) equipped with a blue laser (405 nm). Normal mammary fat pad tissue-derived exosomes were obtained by culturing five mammary fat pads isolated from healthy 4–6-week-old C57BL/6 mice in 3 ml of FBS-free RPMI for 12 h. The final exosome pellet was resuspended in PBS and protein concentration was measured by BCA (Pierce, Thermo Fisher Scientific).

**Proteomics analysis.** Mass spectrometry analyses of exosomes were performed at the Rockefeller University Proteomics Resource Center using 20  $\mu\text{g}$  of exosomal protein. Samples were denatured using 8 M urea, reduced using 10 mM dithiothreitol (DTT), and alkylated using 100 mM iodoacetamide. This was followed by proteolytic digestion with endoproteinase LysC (Wako Chemicals) overnight at room temperature, and subsequent digestion with trypsin (Promega) for 5 h at 37°C. The digestion was quenched with formic acid and resulting peptide mixtures were desalted using in-house made C18 Empore (3M) StAGE tips<sup>47</sup>. Samples were dried and solubilized in the sample loading buffer containing 2% acetonitrile and 2% formic acid. Approximately 3–5  $\mu\text{g}$  of each sample was analysed by reversed phase nano-liquid chromatography–tandem mass spectrometry (LC–MS/MS) (Ultimate 3000 coupled to QExactive, Thermo Scientific). After loading onto the C18 trap column (5  $\mu\text{m}$  beads, Thermo Scientific) at a flow rate of 3  $\mu\text{l min}^{-1}$ , peptides were separated using a 75- $\mu\text{m}$  inner diameter C18 column (3  $\mu\text{m}$  beads, Nikkyo Technos Co.) at a flow rate of 200  $\text{nl min}^{-1}$ , with a gradient increasing from 5% buffer B (0.1% formic acid in acetonitrile)/95% buffer A (0.1% formic acid) to 40% buffer B/60% buffer A, over 140 min. All LC–MS/MS experiments were performed in data-dependent mode. Precursor mass spectra were recorded in a 300–1,400  $m/z$  mass range at 70,000 resolution, and 17,500 resolution for fragment ions (lowest mass:  $m/z$  100). Data were recorded in profile mode. Up to 20 precursors per cycle were selected for fragmentation and dynamic exclusion was set to 45 s. Normalized collision energy was set to 27.

**Semi-quantitative data analysis.** MS/MS spectra were extracted and searched against Uniprot complete human or mouse proteome databases (January 2013) concatenated with common contaminants<sup>48</sup> using Proteome Discoverer 1.4 (Thermo Scientific) and Mascot 2.4 (Matrix Science). All cysteines were considered alkylated with acetamide. N-terminal glutamate to pyroglutamate conversion, oxidation of methionine, and protein N-terminal acetylation were allowed as variable modifications. Data were first searched using fully tryptic constraints. Matched peptides were filtered using a Percolator<sup>49</sup> based 1% false discovery rate (FDR). Spectra not being matched at a FDR of 1% or better were re-searched allowing for semi-tryptic peptides. The average area of the three most abundant

peptides for a matched protein<sup>50</sup> was used to gauge protein amounts within and between samples.

**Label-free quantitative mass spectrometry.** LC–MS/MS data from three technical replicates of six organ-tropic samples were analysed using MaxQuant (version 1.5.0.30) and Perseus software (version 1.5.0.9)<sup>51</sup>, searching against a Uniprot human database (July 2014). Oxidation of methionine and protein N-terminal acetylation were allowed as variable modifications, and cysteine carbamidomethyl was set as a fixed modification. Two missed cleavages were allowed for specificity: trypsin/P. The ‘match between runs’ option was enabled. FDR values at the protein and peptide level were set to 1%. Protein abundance is expressed as LFQ values. Only proteins quantified in at least two out of three replicates in at least one group were retained, and missing values were imputed. A multiple sample ANOVA test was performed and corrected for multiple hypotheses testing using a permutation-based FDR threshold of 0.05.

**Exosome treatment and labelling.** To assess lung, liver and bone exosome distribution, exosomes were injected via the retro-orbital venous sinus, the tail vein or intracardially. Exosome distribution patterns were consistent regardless of the route of injection. For brain distribution, exosomes were only observed in the brain after intracardiac injection. For 24-h exosome treatments, 10  $\mu\text{g}$  of total exosomal protein were injected via the retro-orbital venous sinus, the tail vein, or intracardially in a total volume of 100  $\mu\text{l}$  PBS. For exosome-tracking purposes, purified exosomes were fluorescently labelled using PKH67 (green) or PKH26 (red) membrane dye (Sigma-Aldrich) or FM1-43FX dye (Life Technologies) for the photo-conversion experiment. Labelled exosomes were washed in 20 ml of PBS, collected by ultracentrifugation and resuspended in PBS. When performing peptide blocking experiments, exosomes were incubated with 0.06  $\mu\text{M}$  RGD or HYD-1 (peptide sequence: KIKMVISWKG) peptides for 30 min at 37°C before exosome injection. An average of five random fields was counted per sample at 20 $\times$  magnification, and representative pictures were taken at 40 $\times$  magnification. For education experiments, mice received 10  $\mu\text{g}$  of exosomes retro-orbitally every other day for 3 weeks. To measure exosome uptake by specific cell types, labelled exosomes were injected 24 h before tissue collection and tissues were analysed for exosome-positive cells by immunofluorescence. Pictures were taken at 60 $\times$  magnification. For *in vitro* uptake assays, the membrane of WI-38 cells was labelled with PKH67 dye while 4175-LuT exosomes were labelled with PKH26 dye. Exosomes (10  $\mu\text{g ml}^{-1}$ ) were first incubated with PBS or HYD-1 peptide for 30 min at 37°C, followed by an incubation for 1 h with WI-38 cells at 37°C. Excess exosomes were washed off and pictures were taken by Nikon confocal microscope (Eclipse TE2000U). The amount of exosomes localizing to the lung was analysed by immunofluorescence or using the Odyssey imaging system (LI-COR Biosciences). In brief, NIR dye-labelled exosomes were injected 24 h before tissue collection and tissues were analysed for exosome-positive areas. Whole-lung images were analysed using image J software, quantifying red fluorescence area in arbitrary units.

**Photoconversion and electron microscopy processing.** Cryostat sections prepared at a 15- $\mu\text{m}$  thickness were placed on glass slides and re-fixed in 0.075 M sodium cacodylate, pH 7.4, containing 2.5% glutaraldehyde. For photoconversion, slides were washed twice in 0.1 M sodium cacodylate buffer, pH 7.4. Autofluorescence was quenched using 100 mM  $\text{NH}_4\text{Cl}$  in cacodylate buffer for 45 min. On the basis of optimization experiments, sections were photoconverted for 2 h by incubation in 5.4  $\text{mg ml}^{-1}$  3,3'-diaminobenzidine in 0.1 M sodium cacodylate buffer, pH 7.4, and exposure to the light of an Intensilight C-HGFI 130-W mercury lamp and a 4 $\times$ /0.1 NA objective (Nikon Inverted Microscope Eclipse Ti).

For electron microscopy processing, sections were post-fixed in 1% osmium tetroxide buffer for 15 min on ice. After washing with water, slides were placed in 1% aqueous uranyl acetate for 30 min. Sections were washed with water, dehydrated in a graded series of ethanol concentrations and subsequently in acetone for 10 min at room temperature. Samples were embedded in Eponate. Serial sections were cut at 70 nm in thickness and transferred to formvar-coated slot grids and imaged on a JEOL 100CX at 80 kV with an AMT XR41 digital imaging system.

**Gene expression analysis.** Cell lines were analysed for specific genes using pre-designed TaqMan assays (Applied Biosystems). In brief, RNA was extracted from tissues or cells using the RNeasy kit (Qiagen), and reverse transcribed using Superscript Vilo (Life Technologies). qRT–PCR was performed on a 7500 Fast Real Time PCR System (Applied Biosystems), using TaqMan Universal PCR Master Mix (Applied Biosystems). Relative expression was normalized to  $\beta$ -actin levels.

**Knockdown and overexpression cell preparation.** For shRNA-mediated knockdown of ITGB4 and ITGB5, specific interfering lentiviral vectors containing GFP reporter and puromycin resistance gene cassettes were used. In brief, oligonucleotide 5'-CCGGGAGGGTGTTCATCACCATTGAACTCGAGTTCAATGGTGATGACACCCTCTTTTGTG-3' targeting the 5'-GAGGGTGTTCATCACCATTGAA-3' sequence in the human ITGB4 gene (EntrezGene ID: 3691) or oligonucleotide 5'-CCGGAGCTTGTGTGCCAATGAAATCTCGAGATTTTCATTGGGACAACAGCTTTTTTGTG-3' targeting the 5'-AGCTTGTGTGCCAATGAAAT-3'



sequence in the human *ITGB5* gene (EntrezGene ID: 3693) were cloned into the pLKO.1 vector. As a control, we used the empty pLKO.1 vector. For retrovirus production for integrin overexpression, the pWZL and pBabe vectors systems were used. pWZL-hygro-ITGB<sub>4</sub> and pBabe-puro-ITGB<sub>4</sub> were provided by F. Giancotti. Lentiviral and retroviral particles were packaged using 293T cells. Infected target cells were selected using 500 µg ml<sup>-1</sup> hygromycin B or 2 µg ml<sup>-1</sup> puromycin (Invitrogen).

**Flow cytometry analysis.** Bone marrow was prepared for flow cytometry as previously described<sup>1</sup>. For analysis of lung, tissues were minced and then digested at 37 °C for 20 min with an enzyme cocktail (collagenase A, dispase and DNaseI, Roche Applied Science). Single-cell suspensions were prepared by filtering through a 70-µm strainer and passing through an 18G syringe. Lung fibroblasts were identified by flow cytometry using an anti-mouse rabbit polyclonal S100A4 (1:50, Abcam; ab27957), or SPC (1:100, Santa Cruz; FL-197), revealed by Alexa Fluor 568-conjugated goat anti-rabbit secondary (A-11011, Life Technologies, 1:400). For liver, tissues were mechanically dissociated, and single-cell suspensions were filtered through a 40-µm strainer. Phycoerythrin-conjugated F4/80 (1:100, eBioscience; clone BM8) was used to identify liver macrophages by flow cytometry. Cell fluorescence indicating fluorescently labelled exosome uptake was analysed using a FACSCalibur or a FACSCanto (Beckton Dickinson). FACS data was analysed with FlowJo software (TreeStar Inc.).

**Migration assay.** Twenty-thousand cells were plated in 24-well transwell plates with inserts (8-µm pore size, Corning) and were incubated at 37 °C for 6 h. Cell inserts were fixed with 4% paraformaldehyde (PFA) for 10 min, followed by PBS wash and haematoxylin staining to allow visualization and counting. Nine random fields were counted per well at 20× magnification and the average number of migrated cells per field was calculated.

**Human studies.** Human peripheral blood samples were obtained from control healthy subjects and cancer patients with lung or liver metastasis, or from patients without distant metastasis at Weill Cornell Medical College, University Medical Center Hamburg-Eppendorf, Oslo University Hospital, Memorial Sloan Kettering Cancer Center and University of Nebraska Medical Center, all pathologically confirmed. All individuals provided informed consent for blood donation on approved institutional protocols (WCMC IRB 0604008488 (DL), MSKCC IRB 12-137A (JB)). Plasma or serum exosomes were isolated as previously described<sup>1</sup>. ITGB<sub>4</sub> and ITGA<sub>6</sub> levels in exosomes were measured by ELISA (ABIN417641 and ABIN417609 from Antibodies Online, and LS-F7188 from LifeSpan Biosciences), using 2 µg of exosomes per 100 µl of sample diluent, in duplicate reactions, according to the manufacturer's instructions.

**Mouse studies.** All mouse work was performed in accordance with institutional, IACUC and AAALAS guidelines, by the animal protocol 0709-666A. All animals were monitored for abnormal tissue growth or ill effects according to AAALAS guidelines and euthanized if excessive deterioration of animal health was observed. No statistical method was used to pre-determine sample size. No method of randomization was used to allocate animals to experimental groups. The investigators were not blinded to allocation during experiments and outcome assessment. Mice that died before the predetermined end of the experiment were excluded from the analysis. In none of the experiments did tumours exceed the maximum volume allowed according to our IACUC protocol, specifically 2 cm<sup>3</sup>. For exosome localization, education and tumour implantation experiments for mouse cell lines, 6-week-old C57BL/6 *Mus musculus* females purchased from Jackson labs were used. For exosome localization, education and tumour implantation experiments for human cell lines, 6–8-week-old NCr nude (*NCRNU-F* sp./sp.) females purchased from Taconic were used. For lung metastasis studies using organotrophic lines, 6–8-week-old nude female mice were pre-educated with exosomes for 3 weeks followed by tail vein injection of 2 × 10<sup>5</sup> or intracardiac injection of 1 × 10<sup>5</sup> luciferase-positive cancer cells resuspended in 100 µl PBS. Four weeks after intracardiac injection and eight weeks after tail vein injection, lung metastasis was measured using the IVIS 200 bioluminescence imaging system (Xenogen, Caliper Life Sciences), and tissues were cut in 6-µm sections and stained with haematoxylin and eosin for histology. To analyse the role of exosome education in tumour metastasis, 6–8-week-old C57BL/6 female mice pre-educated with pancreatic cancer-derived exosomes were injected intraperitoneally with 1 × 10<sup>6</sup> Pan02 mCherry cells resuspended in 30 µl of Matrigel (Corning). One or twenty-one days later, mice were euthanized, and livers were analysed for metastatic lesions by measuring liver weight.

To follow the levels of tumour-derived exosomes in plasma of tumour-bearing mice, 1 × 10<sup>6</sup> 4175 lung-tropic cells were injected in the mammary fat pad of nude mice. Mouse blood (250 µl) was drawn from the retro-orbital sinus when tumour size was over 800 mm<sup>3</sup>, followed by tumour resection. One week after the tumour was resected, mice were analysed by bioluminescence IVIS imaging for luciferase activity and separated into two groups: recurrence/tumour-free and recurrent tumours. Mouse blood was drawn and the plasma of mice within the same group was pooled for exosome isolation. Western blot analysis with anti-human ITGB<sub>4</sub> antibodies was used to detect tumour-derived exosomes.

To assess exosome-induced vascular leakiness, 10 µg of total exosome protein were injected by retro-orbital injection. Then 20 h after exosome treatment, mice were injected with 2 mg of Texas Red-lysine fixable dextran 70,000 MW (Invitrogen) via retro-orbital injection. One hour after dextran injection, mice were euthanized and perfused with PBS. Lungs were dissected and fixed in a mix of 2% PFA and 20% sucrose overnight, then embedded in Tissue-tek O.C.T. embedding compound (Electron Microscopy Sciences) and frozen in a dry-ice/ethanol bath. O.C.T. blocks were sectioned and stained for DAPI, pictures were taken using a Nikon confocal microscope (Eclipse TE2000U). Images were analysed using image J software, quantifying red fluorescence area in arbitrary units.

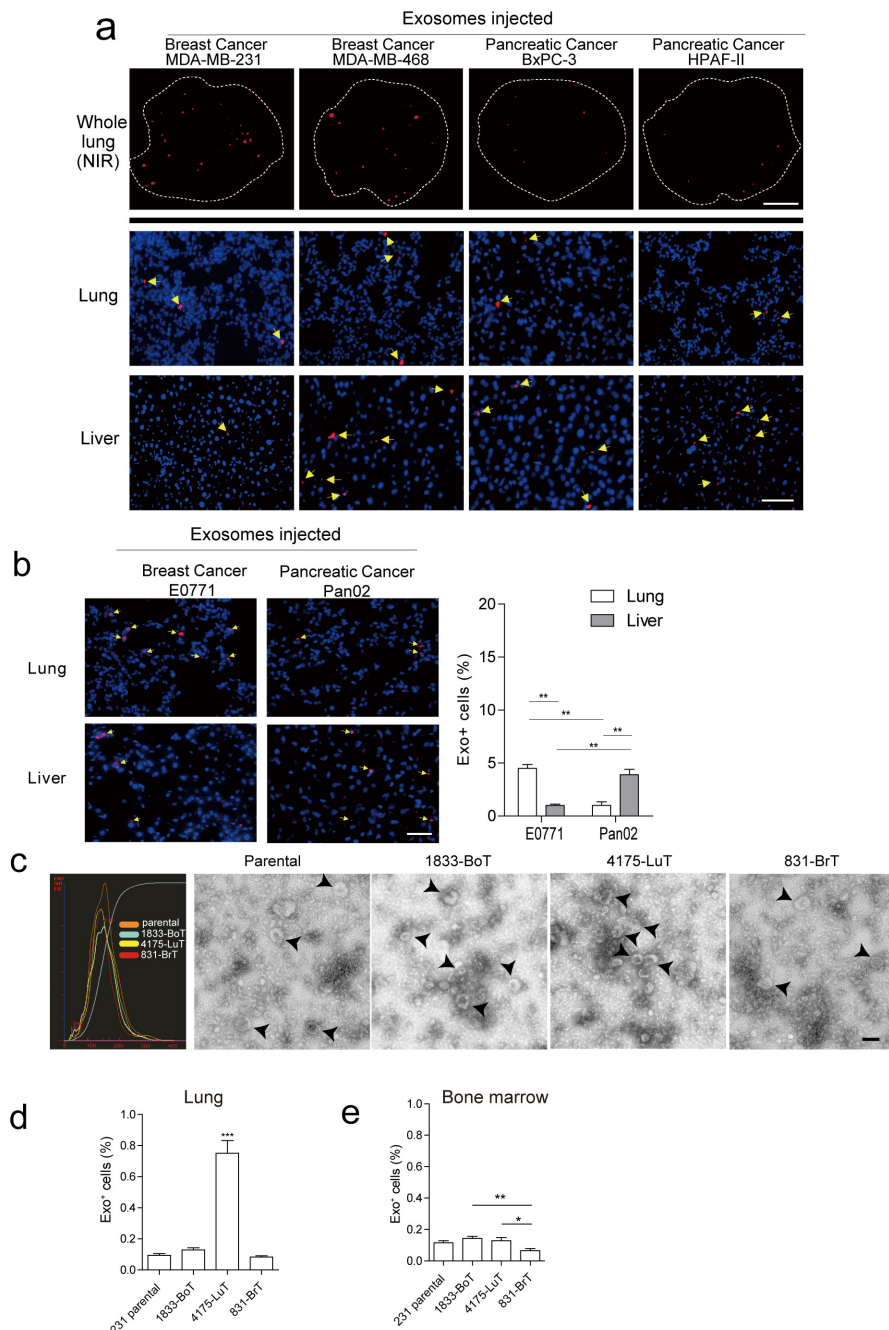
**Tissue processing and immunofluorescence.** For histological analysis, tissues were dissected and fixed in a mix of 2% PFA and 20% sucrose in PBS overnight, then embedded in Tissue-tek O.C.T. embedding compound. Blocks were frozen in a dry-ice/ethanol bath. For immunofluorescence, 6 µm O.C.T. tissue cryosections were stained with antibodies against F4/80 (1:100, eBioscience; BM8), fibronectin (1:50, Santa Cruz; IST-9), S100A4 (1:100, Abcam; ab27957), SPC (1:100, Santa Cruz; FL-197), laminin (1:50, abcam; ab11575), CD31 (1:100, Santa Cruz; MEC 13.3), EpCAM (1:50, Santa Cruz; HEA125). Secondary antibodies conjugated to Alexa Fluor 488 or 549 were used (A-11001 and A-11007, Life Technologies). Fluorescent images were obtained using a Nikon confocal microscope (Eclipse TE2000U) and analysed using Nikon software (EZ-C1 3.6).

**Western blot analysis.** Exosomes or cells were lysed with RIPA buffer containing a complete protease inhibitor tablet (Roche). Lysates were cleared by centrifugation at 14,000g for 20 min. Supernatant fractions were used for western blot. Samples were separated on a Novex 4–12% Bis-Tris Plus Gel (Life Technologies), and transferred onto a PVDF membrane (Millipore). Membranes were processed for Ponceau red staining followed by 1 h blocking and primary antibody incubation. The antibodies against the following proteins were used for western blot analysis: ITGB<sub>1</sub> (1:1,000, Cell Signaling; 4706), ITGB<sub>4</sub> (1:500, Cell Signaling; 4707), ITGA<sub>6</sub> (1:1,000, Cell Signaling; 3750), ITGA<sub>2</sub> (1:10,000, abcam; ab133557), ITGA<sub>3</sub> (1:1,000, abcam; ab190731), ITGA<sub>7</sub> (1:500, abcam; ab117611), ITGB<sub>5</sub> (1:500, Cell Signaling; 4708), ITGB<sub>3</sub> (1:500, Millipore; AB2984) Alix (1:1,000, Cell Signaling; 3A9), and GAPDH (1:10,000, Cell Signaling; 14C10). Anti-rabbit IgG, horseradish peroxidase (HRP)-linked antibody (1:3,000, Cell Signaling; 7074) and anti-mouse IgG, HRP-linked antibody (1:3,000, Cell Signaling; 7076) were used as secondary antibodies.

**In situ protein expression analysis (in-cell western assay, LI-COR).** Cells were plated in a 96-well plate and treated with 10 µg ml<sup>-1</sup> exosomes for 2 h and then processed according to the protocol provided by the manufacturer. In brief, cells were fixed with 4% PFA and washed with 0.1% TritonX-100/PBS. Cells were then blocked using Odyssey blocking buffer for 1 h and stained overnight at 4 °C with primary antibody in Odyssey blocking buffer containing 0.1% Tween-20. The next day cells were washed again and incubated with LI-COR secondary antibodies for 1 h at room temperature followed by fluorescent imaging using Odyssey. Antibodies against the following proteins were used: Src (1:100, Cell Signaling; 2109), p-Src (1:100, Cell Signaling; 2101), AKT (1:100, Cell Signaling; 9272), p-AKT (1:100, Cell Signaling; 9271), p38 (1:100, Cell Signaling; 9212), p-p38 (1:100, Cell Signaling; 9211), NF-κB (1:100, Cell Signaling; 3034), p-NF-κB (1:100, Cell Signaling; 3033), NFAT (1:100, Thermo Scientific; PA1-023), ILK (1:100, abcam; ab52480), FAK (1:100, abcam; ab40794) and GAPDH (1:100, Cell Signaling; 14C10). IRDye 800CW anti-rabbit IgG (1:800, LI-COR) were used as secondary antibodies.

**Statistical analysis.** Error bars in graphical data represent mean ± s.e.m. Mouse experiments were performed in duplicate or triplicate, using 3–6 mice per treatment group. Statistical significance was determined using a two-tailed Student's *t*-test and one-way ANOVA, in which *P* values of *P* < 0.05 were considered statistically significant. Variance was similar between the groups that were statistically compared.

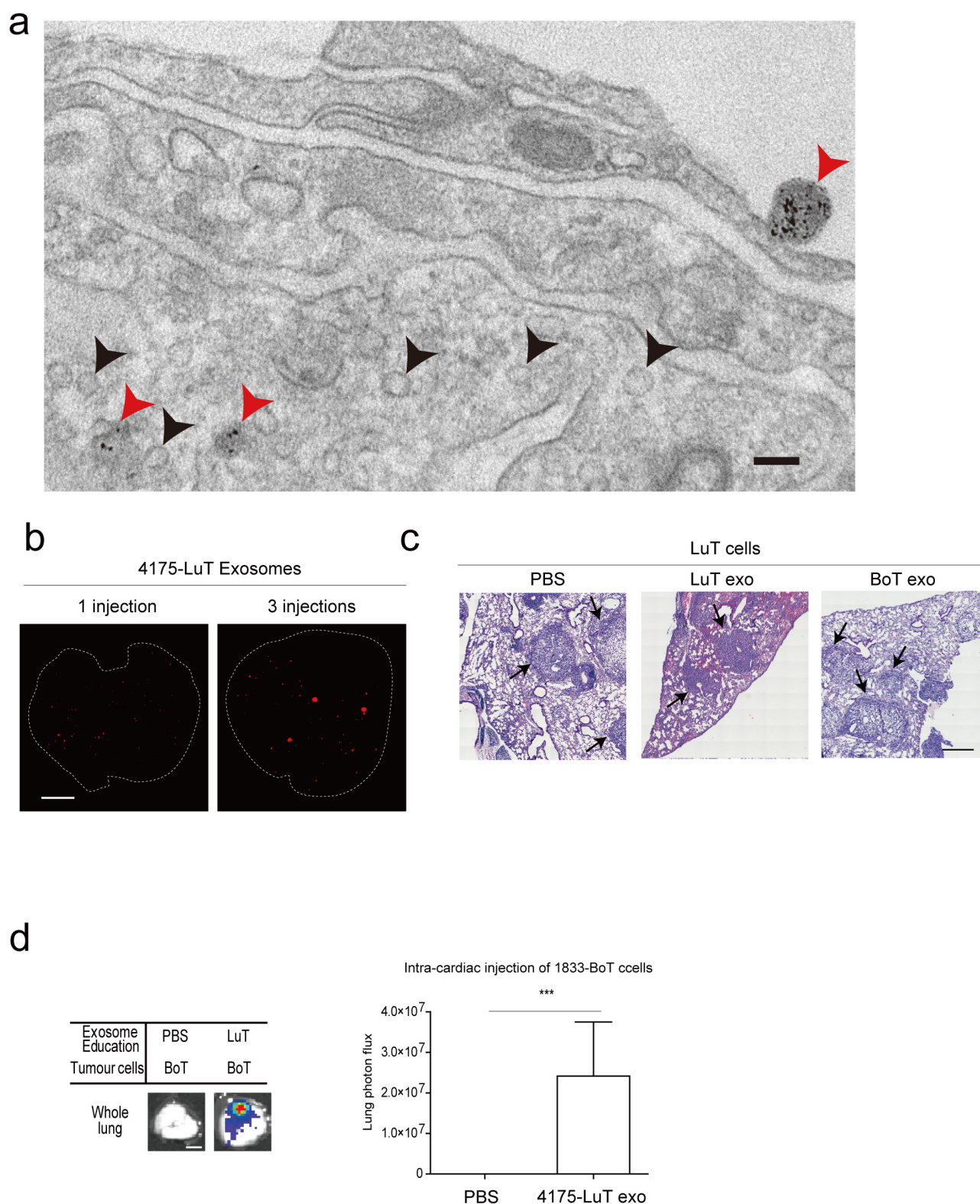
46. Peinado, H. *et al.* Melanoma exosomes educate bone marrow progenitor cells toward a pro-metastatic phenotype through MET. *Nature Med.* **18**, 883–891 (2012).
47. Rappsilber, J., Ishihama, Y. & Mann, M. Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal. Chem.* **75**, 663–670 (2003).
48. Bunkenborg, J., Garcia, G. E., Paz, M. I., Andersen, J. S. & Molina, H. The minotaur proteome: avoiding cross-species identifications deriving from bovine serum in cell culture models. *Proteomics* **10**, 3040–3044 (2010).
49. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & Mac Coss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods* **4**, 923–925 (2007).
50. Silva, J. C., Gorenstein, M. V., Li, G. Z., Vissers, J. P. & Geromanos, S. J. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol. Cell. Proteomics* **5**, 144–156 (2006).
51. Cox, J. *et al.* Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* **13**, 2513–2526 (2014).



**Extended Data Figure 1 | Characterization of organotropic exosome properties and biodistribution.** **a**, Human cancer exosome biodistribution in lung and liver. Exosomes (10 µg) derived from each cell line were labelled with lipophilic PKH26 dye (red) and injected retro-orbitally into nude mice 24 h before culling. Top, representative NIR whole-lung image by Odyssey imaging ( $n = 3$ ). Middle and bottom, represent exosome biodistribution in the lung and liver as determined by immunofluorescence microscopy. Arrows indicate exosome foci ( $n = 3$ , three independent experiments). **b**, Biodistribution of exosomes isolated from mouse cell lines E0771 and Pan02. Mouse exosome biodistribution in the lung and liver was determined by immunofluorescence microscopy. Exosomes (10 µg) derived from each cell line were labelled with lipophilic PKH26 dye (red) and injected retro-orbitally into nude mice 24 h before culling. Top, lung at 40× magnification. Bottom, liver at 40× magnification. Arrows indicate exosome foci. Graph represents the quantification of exosome distribution by counting exosome-positive cells. An average of five random fields per sample were counted at 20× magnification (three independent experiments, each with  $n = 3$ ).  $**P < 0.01$  by two-tailed Student's  $t$ -test. **c**, Analysis of organotropic cell-derived exosomes. MDA-MB-231 organotropic cell-line-derived exosomes

were analysed for size distribution by NanoSight and phenotype (purity and shape) by electron microscopy; black arrows indicate representative exosomes. Technical triplicates were analysed, at least 10 images per sample. **d**, Flow cytometric analysis of exosome<sup>+</sup> cells in lung. Exosomes (10 µg) derived from MDA-MB-231 organotropic cell lines were labelled with lipophilic PKH67 dye (green) and injected retro-orbitally into nude mice 24 h before culling. FITC-channel-positive cells were acquired on a FACS Calibur, and the percentage of exosome-positive cells was quantified (representing data pooled from two independent experiments, a total of  $n = 12$ ).  $**P < 0.001$  by one-way ANOVA. **e**, Flow cytometric analysis of exosome-positive cells in the bone marrow. Exosomes (10 µg) derived from MDA-MB-231 organotropic cell lines were labelled with lipophilic PKH67 dye (green) and injected retro-orbitally into nude mice 24 h before culling. FITC-channel-positive cells were acquired on a FACS Calibur, and the percentage of exosome-positive cells was quantified (representative data pooled from two independent experiments, a total of  $n = 6$ ).  $**P < 0.01$  and  $*P < 0.05$  by one-way ANOVA for the 831-BrT to 1833-BoT and 4175-LuT comparisons, respectively. Data are mean  $\pm$  s.e.m. Scale bars, 5 mm (**a**, top), 50 µm (**a**, middle and bottom, **b**) and 100 nm (**c**).



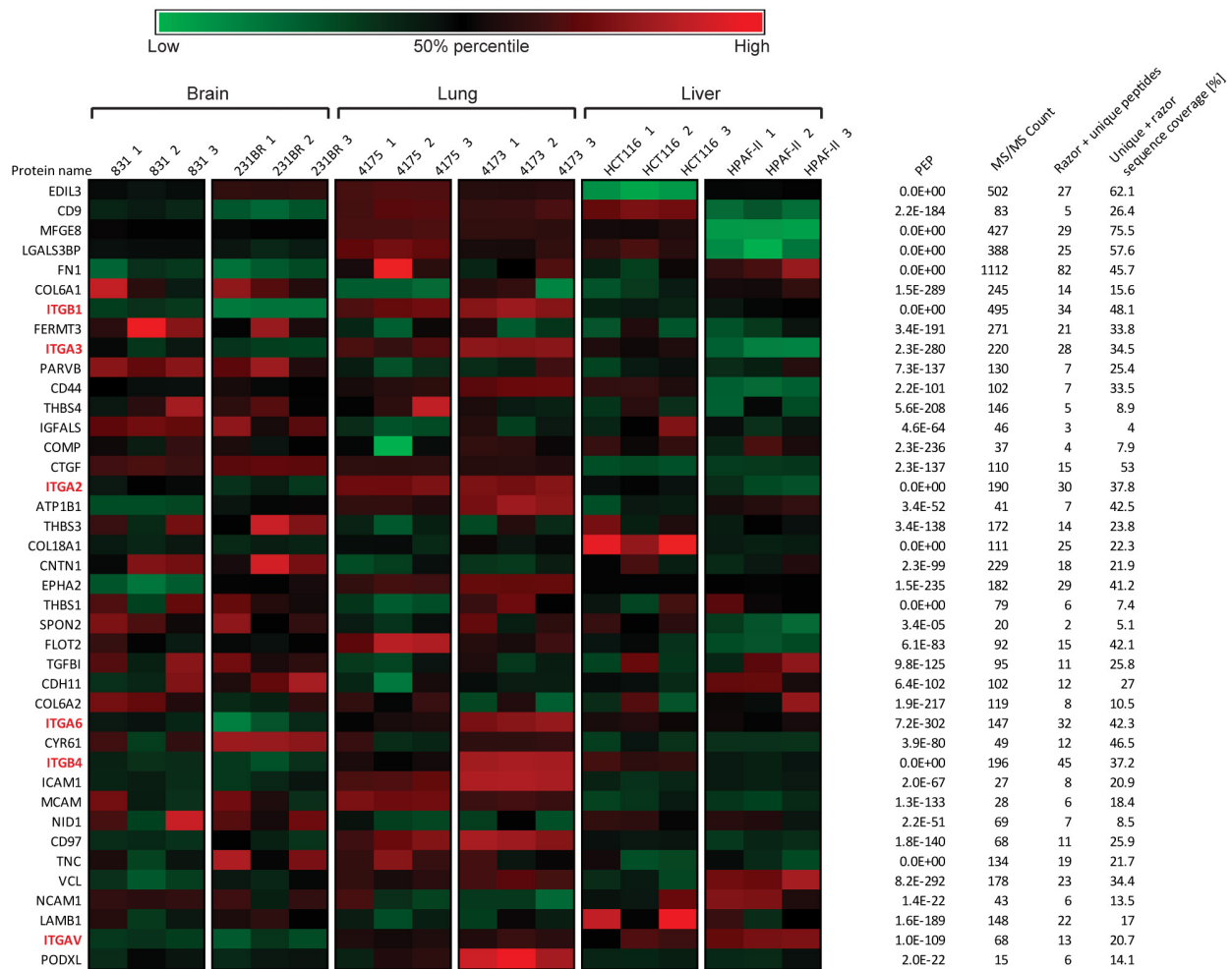


**Extended Data Figure 2 | 4175-LuT cell-derived exosomes localize to lung and dictate future metastatic sites.** **a**, Electron microscopy imaging of FM1-43-labelled 4175-LuT exosomes. Red arrows, FM1-43-positive exogenous exosomes; black arrows, endogenous exosomes. Two mice were tested, images were taken for several sections from each organ ( $n = 30$  images in total). **b**, Representative NIR imaging of lung whole mount after daily exosome injections. Exosomes ( $10 \mu\text{g}$ ) derived from 4175-LuT cells were injected daily for three consecutive days via the retro-orbital sinus and the whole lung was imaged by Odyssey imaging ( $n = 4$ ). **c**, Representative haematoxylin/eosin staining of the lung from Fig. 1f

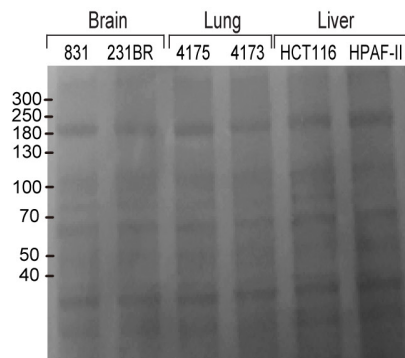
at  $20\times$  magnification;  $n = 5$  for all, except for LuT exo/LuT cells, in which  $n = 4$ ; data representative of two independent experiments. Arrows indicate lung metastasis. **d**, Analysis of 1833-BoT cell metastasis to the lung, after 3 weeks of continuous treatment with PBS or 4175-LuT exosomes, followed by intracardiac injection of  $1 \times 10^5$  tumour cells. Mice were injected retro-orbitally with exosomes every other day for 3 weeks, before tumour cell injection. Quantitative bioluminescence imaging of luciferase activity by IVIS imaging. Metastasis was quantified 3 weeks after tumour cell injection ( $n = 4$ ). Scale bars,  $100 \text{ nm}$  (**a**),  $5 \text{ mm}$  (**b**, **d**) and  $500 \mu\text{m}$  (**c**). Data are mean  $\pm$  s.e.m. \*\*\* $P < 0.001$  by two-tailed Student's  $t$ -test.



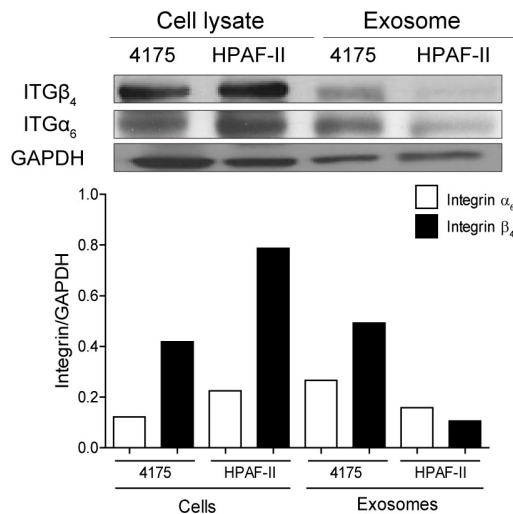
a



b



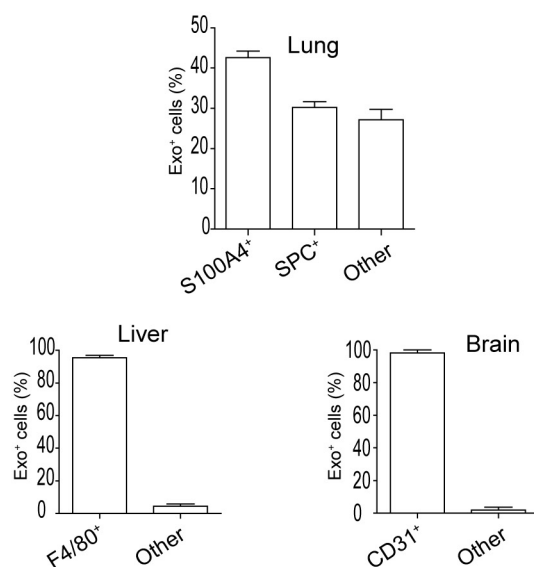
c



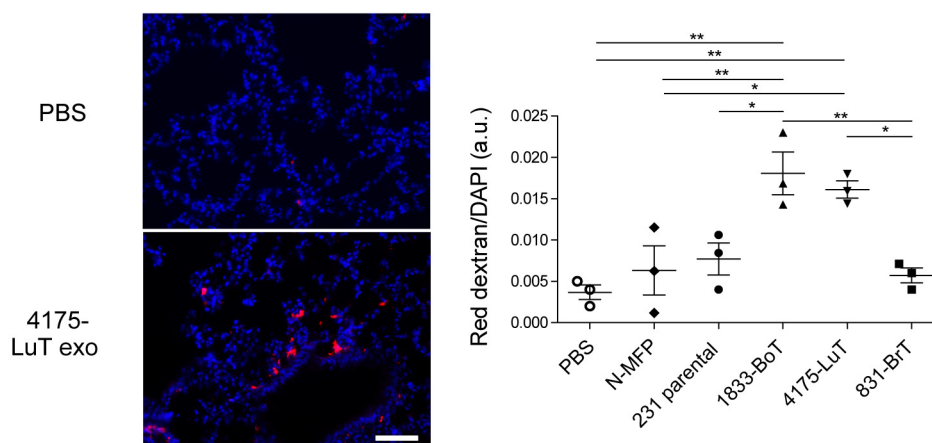
**Extended Data Figure 3 | Characterization of organotropic exosome protein cargo.** **a**, Top 40 adhesion molecules packaged in exosomes isolated from organotropic cell lines. Heat map of adhesion molecule signals based on Z-scored LFQ values obtained from quantitative mass spectrometry analysis. PEP (posterior error probability), MS/MS count is a number of fragmentation spectra (spectral counting), Razor + unique peptides refers to the number of peptides, and sequence coverage refers to percentage of peptide counts identified. **b**, Ponceau staining of exosome lysates isolated

from organotropic cell lines. Representative Ponceau staining of total protein from the organotropic cell-line-derived exosomes. Exosomal protein (10  $\mu$ g) was loaded in each well ( $n = 2$ , three independent experiments). **c**, Western blot analysis comparison of ITG $\alpha_6$  and ITG $\beta_4$  levels in cell lysates versus exosomes derived from organotropic breast cancer and pancreatic cancer cell lines. Graph represents the relative ratios of integrin to GAPDH signals as determined by densitometry. For western blot source data, see Supplementary Fig. 1i–k.

a



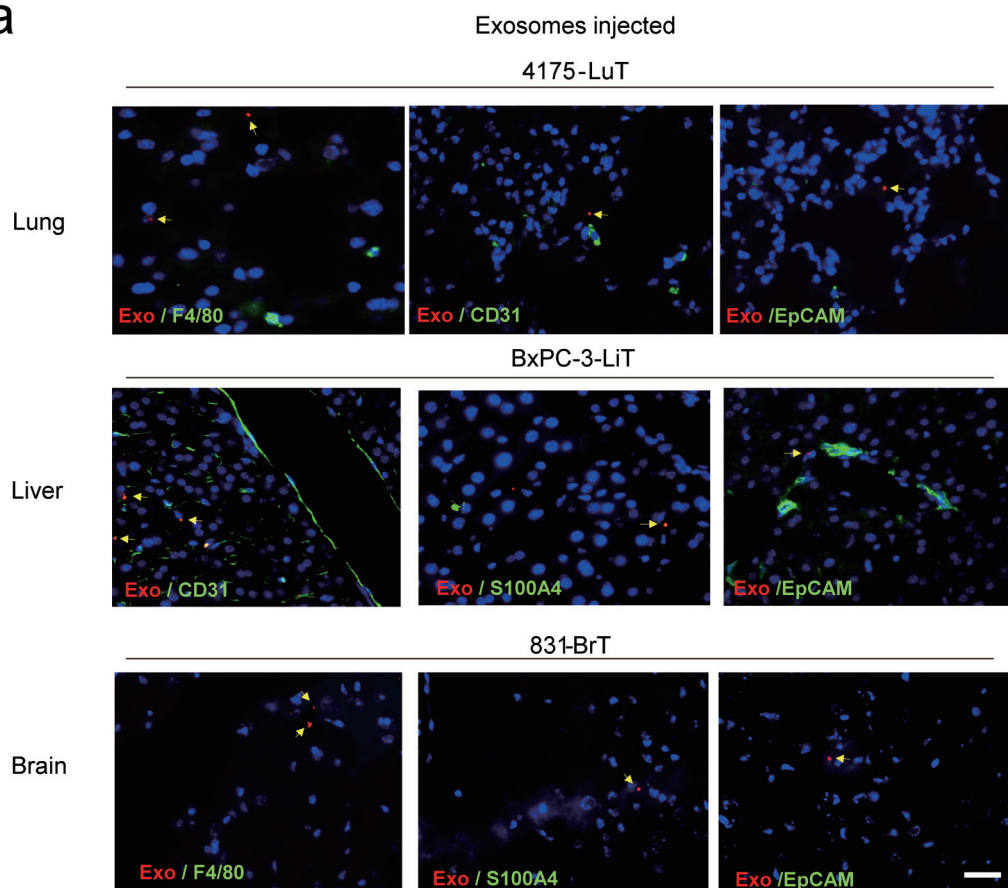
b



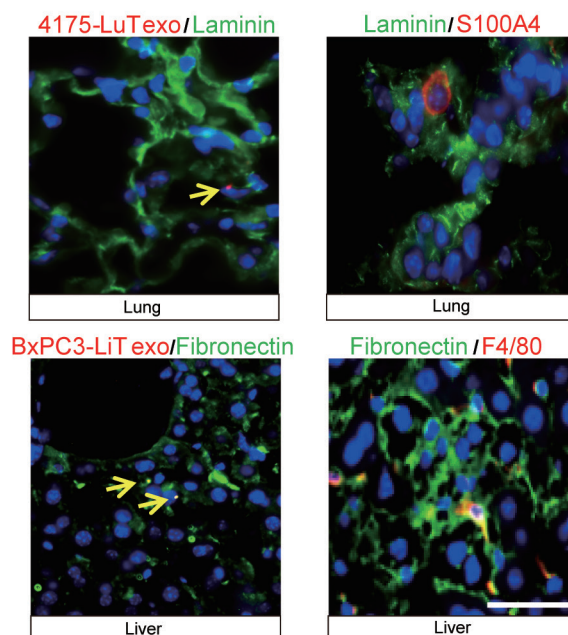
**Extended Data Figure 4 | Functional characterization of organotrophic exosomes.** **a**, Quantification of organotrophic exosome uptake by target cells *in vivo*. Top graph, flow cytometric quantification of the frequency of 4175-LuT exosome-positive fibroblasts and epithelial cells ( $n = 4$ ). Left bottom graph, flow cytometric quantification of the frequency of BxPC-3 exosome-positive macrophages ( $n = 3$ ). Right bottom graph, quantification of the frequency of 831-BrT exosome-positive endothelial cells by immunofluorescence microscopy ( $n = 5$ ). **b**, Organotrophic cell-line-derived exosomes induce vascular leakiness in the lung. Leakiness in the

lung 24 h after retro-orbital injection of 10 μg of normal mammary fat pad or MDA-MB-231 organotrophic cell-line-derived exosomes was quantified by imaging the presence of fluorescent dextran (red) outside of blood vessels, in the lung parenchyma. Left top panel, 40× magnification of representative lung image after PBS injection. Left bottom panel, representative lung image after 4175-LuT exosome injection. Scale bar, 50 μm. Right graph depicts the quantification of five random areas at 20× magnification in arbitrary units (data representative of two independent experiments;  $n = 3$ ). Data are mean ± s.e.m. \* $P < 0.05$ ; \*\* $P < 0.01$  by one-way ANOVA.

a



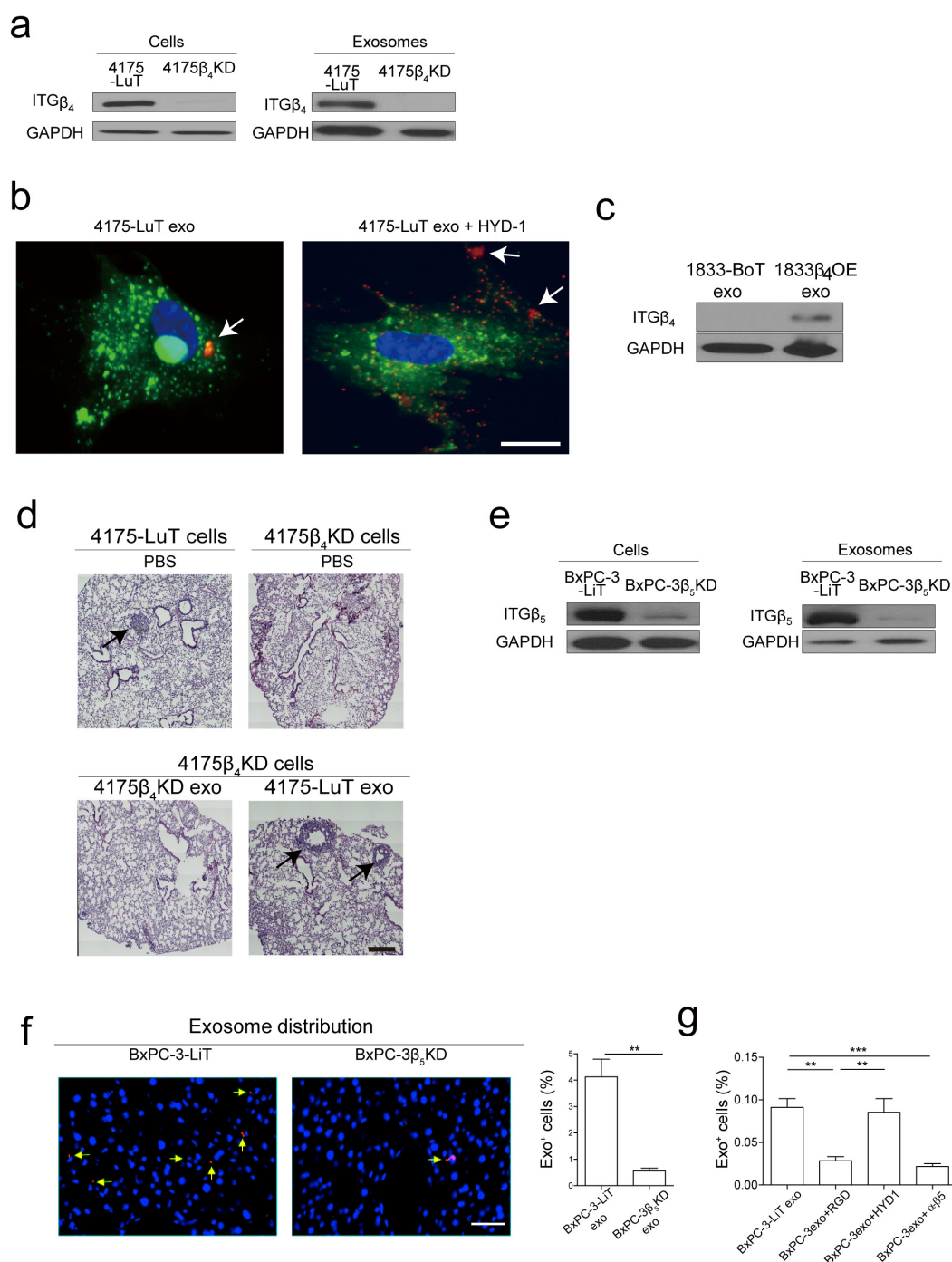
b



**Extended Data Figure 5 | Exosome co-localization with specific cell types within target tissues.** **a**, Immunofluorescence analysis of resident cells in lung, liver and brain after labelled exosome injection. Analysis of exosome (red) co-staining with markers (green) for tissue-specific stromal cell types. Top, representative images of immunofluorescence microscopy of 4175-LuT exosome co-staining with F4/80, CD31 and EpCAM. Middle, liver sections from mice injected with BxPC-3-LiT-derived exosomes were co-stained with CD31, S100A4 and EpCAM. Bottom, brain sections from mice injected with 831-BrT exosome were co-stained with F4/80, S100A4 and EpCAM

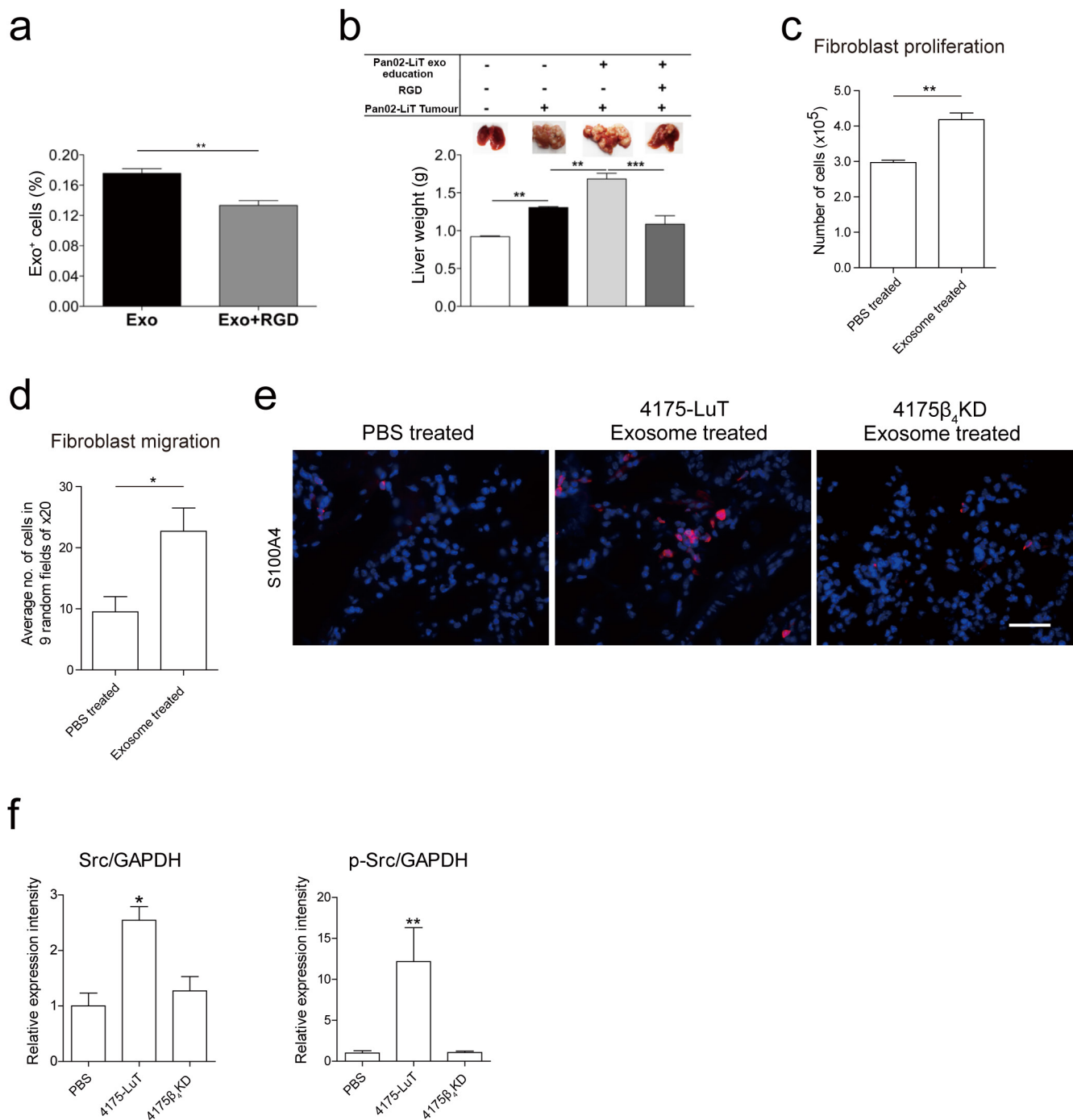
( $n = 3$  per experiment for two independent experiments). **b**, Exosome biodistribution and co-localization with extracellular matrix proteins. Left top, representative immunofluorescence microscopy images of lung tissue, depicting 4175-LuT exosome (red) co-staining with laminin (green). Right top, laminin (green) co-staining with S100A4 (red). Left bottom, representative immunofluorescence microscopy of liver tissue co-stained for fibronectin (green) and BxPC-3-LiT exosomes (red). Right bottom, fibronectin (green) co-staining with F4/80 (red) ( $n = 3$ , two independent experiments). Scale bars, 30  $\mu$ m.





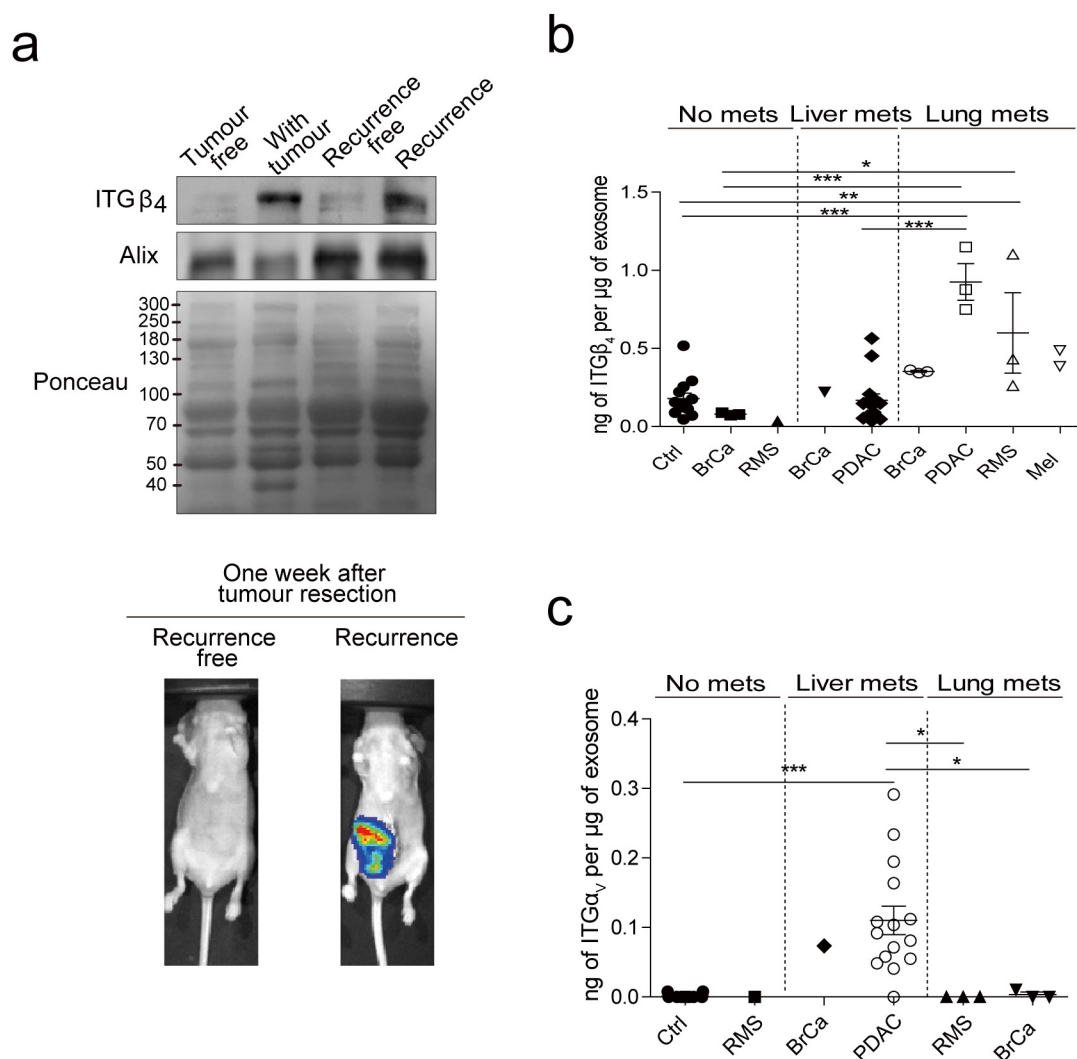
**Extended Data Figure 6 | ITGs functionally regulate organotropic exosome uptake and exosome-mediated metastasis.** **a**, Representative western blot analysis of integrin expression in 4175-LuT and 4175 $\beta_4$ KD cells and exosomes (representative of three independent experiments). For western blot source data, see Supplementary Fig. 11. **b**, *In vitro* uptake of 4175-LuT exosomes by WI-38 lung fibroblasts. The WI-38 cell membrane was labelled with PKH67 green dye and 4175-LuT exosomes were labelled with PKH26 red dye. Exosomes ( $10 \mu\text{g ml}^{-1}$ ) were first incubated with PBS or HYD-1 peptide for 30 min at  $37^\circ\text{C}$ , followed by 1-h incubation with WI-38 cells at  $37^\circ\text{C}$ . Excess exosomes were washed and cells were imaged ( $n = 4$  for two independent experiments). **c**, Representative western blot of ITG $\beta_4$  expression in exosomes isolated from wild-type or ITG $\beta_4$ -overexpressing 1833-BoT cells (representative of two independent experiments). For western blot source data, see Supplementary Fig. 1m. **d**, Representative haematoxylin/eosin staining of lungs from Fig. 3e. Arrows indicate lung metastasis;  $n = 6$ , data representative of two independent experiments. **e**, Representative western blot analysis of integrin expression in BxPC-3-LiT and BxPC-3 $\beta_5$ KD

cells and exosomes. For western blot source data, see Supplementary Fig. 1n. **f**, Immunofluorescence analysis of BxPC-3-LiT control and BxPC-3 $\beta_5$ KD-derived exosome biodistribution in the liver. Exosomes ( $10 \mu\text{g}$ ) isolated from each cell line were labelled with lipophilic PKH26 dye (red) and injected retro-orbitally into nude mice 24 h before culling. Left,  $40\times$  magnification. Arrows indicate exosome foci. Scale bar,  $50 \mu\text{m}$ . Right, quantification of exosome distribution by exosome-positive cells. An average of five random fields were counted at  $20\times$  magnification (data representative of two independent experiments;  $n = 3$ ). **g**, Flow cytometry analysis of exosome-positive cells in the liver 24 h after exosome injection. Labelled BxPC-3-LiT exosomes ( $5 \mu\text{g}$ ) per mouse were incubated with PBS, RGD, HYD-1 or ITG $\alpha_v\beta_5$  antibody for 30 min at  $37^\circ\text{C}$  before retro-orbital injection into nude mice. Livers were collected and analysed for exosome-positive cells by flow cytometry 24 h after injection ( $n = 4$ , except for the ITG $\alpha_v\beta_5$  antibody group, in which  $n = 5$ ). Scale bars,  $10 \mu\text{m}$  (**b**),  $500 \mu\text{m}$  (**d**) and  $500 \mu\text{m}$  (**f**).  $**P < 0.01$ ,  $***P < 0.001$  by two-tailed Student's *t*-test (**f**) and one-way ANOVA (**g**). Data are mean  $\pm$  s.e.m.



**Extended Data Figure 7 | Functional contribution of exosomes to metastasis.** **a**, Microscopic analysis of exosome-positive cells in the livers of mice injected with liver metastatic Pan02-LiT-derived exosomes. Before injection, Pan02-LiT exosomes were pre-incubated with RGD peptide for 30 min at 37 °C. Pan02-LiT exosomes (10 μg) were labelled with lipophilic PKH67 green dye and injected retro-orbitally into C57BL/6 mice 24 h before culling. Livers were digested and exosome-positive cells were quantified by flow cytometry ( $n = 3$ ). **b**, Analysis of Pan02-LiT liver metastasis after 3 weeks of continuous treatment with PBS, Pan02-LiT-derived exosomes, or Pan02-LiT-derived exosomes pre-incubated with RGD peptide for 30 min at 37 °C. Pan02-LiT cells were injected intrasplenically. Mice were injected retro-orbitally with 5 μg exosome every other day for 3 weeks. Top, representative liver images showing metastasis taken at culling. Bottom, liver weight quantification ( $n = 4$  except for the control and peptide group for which  $n = 3$  of one experiment). **c**, Functional analysis of lung fibroblasts educated with 4175-LuT-derived exosomes. Proliferation of lung

fibroblasts educated with exosomes every other day for 2 weeks. Three days after cells were plated at equal density, cell numbers were counted using a haemocytometer ( $n = 3$ ; three independent experiments). **d**, Migration of lung fibroblasts educated with exosomes every other day for 2 weeks was measured as follows. Fibroblasts were plated in 24-well transwell chamber inserts, and after 6 h the number of cells that migrated was counted using haematoxylin staining. Nine random fields were counted at 20× magnification and the average number of cells per field was calculated (total of  $n = 4$  from two independent experiments). **e**, Representative image of the lung stained for S100A4. Mice were treated every other day with PBS, 4175-LuT or 4175β<sub>4</sub>KD exosomes for 3 weeks. Scale bar, 50 μm;  $n = 4$  mice. **f**, *In situ* (in-cell western) protein expression analysis of WI-38 fibroblasts treated with PBS, 4175-LuT or 4175ITGβ<sub>4</sub>KD exosomes. Relative expression levels of Src and phosphorylated (p-) Src ( $n = 3$ , three independent experiments). Data are mean ± s.e.m. \* $P < 0.05$ ; \*\* $P < 0.01$ ; \*\*\* $P < 0.001$  by two-tailed Student's *t*-test (**a**, **c**, **d**) and one-way ANOVA (**b**, **f**).



**Extended Data Figure 8 | Exosomal integrin expression as a potential metastatic site biomarker.** **a**, Exosomal ITG $\beta_4$  levels in the plasma of mice bearing orthotopic 4175-LuT tumours, as a function of tumour progression. Blood plasma was collected for exosome isolation 6 weeks after intra-mammary fat pad tumour injection, then again 1 week after tumour resection, from mice that were deemed to be either free of tumour or presenting with recurring tumours based on IVIS bioluminescence imaging ( $n = 5$  were pooled for each group, based on one experiment). For western blot source data, see Supplementary Fig. 10. **b**, Exosomal ITG $\beta_4$  in healthy control subjects (Ctrl) ( $n = 13$ ); patients with breast cancer (BrCa)

and no metastasis ( $n = 3$ ), liver metastasis ( $n = 1$ ), or lung metastasis ( $n = 3$ ); patients with rhabdomyosarcoma (RMS) and no metastasis ( $n = 1$ ) or lung metastasis ( $n = 3$ ); patients with pancreatic cancer (PDAC) with liver metastasis ( $n = 14$ ) and lung metastasis ( $n = 3$ ); and patients with melanoma (Mel) with lung metastasis ( $n = 2$ ). **c**, Exosomal ITG $\alpha_v$  in healthy control subjects ( $n = 13$ ); patients with rhabdomyosarcoma and no metastasis ( $n = 1$ ) or lung metastasis ( $n = 3$ ); patients with breast cancer and lung metastasis ( $n = 3$ ) or liver metastasis ( $n = 1$ ); and patients with pancreatic cancer and liver metastasis ( $n = 15$ ). Data are mean  $\pm$  s.e.m. \* $P < 0.05$ ; \*\*\* $P < 0.001$  by one-way ANOVA.



Extended Data Table 1 | Integrin expression in human exosomes in multiple organotropic tumour models

Sites of metastasis	None		Bone	Brain				Lung										Liver									
Cell type	Lung Fibroblast	Mammary Epithelial	Breast Cancer			Melanoma		Breast Cancer			Osteo-sarcoma	Rhabdomyosarcoma		Wilms' Tumor		Melanoma	Uveal melanoma	Colorectal Cancer			Pancreatic Cancer				Gastric Cancer		
Cell line	WI-38	MCF10A	1833	831	231BR	131/4-5B2	SB1B	4173	4175	4180	143B	RD	CT10	CCG 9911	CLS1	131f/8-2L	Primary culture	HCT116	HT29	SW620	BXPC-3	HPAF-II	MiaPaca-2	PANC-1	SNU1	SNU16	
ITGα <sub>1</sub>							+	+		+							+			+	+						
ITGα <sub>2</sub>			+	+	+	+	+	+	+	+	+			+		+	+	+	+	+	+	+	+				
ITGα <sub>2b</sub>							+														+						
ITGα <sub>3</sub>		+	+	+	+			+	+	+	+						+	+	+		+	+	+				
ITGα <sub>4</sub>						+					+	+		+	+	+	+										
ITGα <sub>5</sub>							+	+	+	+		+	+		+								+	+	+		
ITGα <sub>6</sub>				+	+		+	+	+	+	+	+	+	+	+	+	+			+	+	+	+	+	+	+	
ITGα <sub>11</sub>										+																	
ITGα <sub>v</sub>				+		+	+	+	+	+	+	+	+		+	+	+	+	+	+	+	+	+	+	+	+	
ITGβ <sub>1</sub>		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
ITGβ <sub>3</sub>				+	+	+	+	+		+	+	+	+	+	+	+	+	+	+	+	+	+	+				
ITGβ <sub>4</sub>							+	+	+	+								+	+	+	+	+	+			+	
ITGβ <sub>5</sub>	+			+			+	+	+	+		+	+		+		+	+	+	+	+	+	+	+	+	+	
ITGβ <sub>6</sub>																					+	+	+				

Proteomic analysis of integrins in exosomes. '+' indicates positive for integrin expression by qualitative mass spectrometry.

Extended Data Table 2 | Integrin expression in human and mouse cell-line-derived exosomes

Human			Murine		
Sites of metastasis	Majority to lung	Lung and liver	Sites of metastasis	Lung	Liver
Cell type	Breast cancer		Cell type	Breast cancer	Pancreatic cancer
Cell line	MDA-MB-231	MDA-MB-468	Cell line	E0771	Pan02
<b>ITG<math>\alpha</math><sub>1</sub></b>			<b>ITG<math>\alpha</math><sub>1</sub></b>		
<b>ITG<math>\alpha</math><sub>2</sub></b>	+	+	<b>ITG<math>\alpha</math><sub>2</sub></b>		
<b>ITG<math>\alpha</math><sub>2b</sub></b>			<b>ITG<math>\alpha</math><sub>2b</sub></b>		+
<b>ITG<math>\alpha</math><sub>3</sub></b>	+	+	<b>ITG<math>\alpha</math><sub>3</sub></b>	+	+
<b>ITG<math>\alpha</math><sub>4</sub></b>			<b>ITG<math>\alpha</math><sub>4</sub></b>		
<b>ITG<math>\alpha</math><sub>5</sub></b>		+	<b>ITG<math>\alpha</math><sub>5</sub></b>	+	+
<b>ITG<math>\alpha</math><sub>6</sub></b>	+	+	<b>ITG<math>\alpha</math><sub>6</sub></b>	+	+
<b>ITG<math>\alpha</math><sub>v</sub></b>		+	<b>ITG<math>\alpha</math><sub>v</sub></b>	+	+
<b>ITG<math>\beta</math><sub>1</sub></b>	+	+	<b>ITG<math>\beta</math><sub>1</sub></b>	+	+
<b>ITG<math>\beta</math><sub>3</sub></b>	+	+	<b>ITG<math>\beta</math><sub>3</sub></b>	+	+
<b>ITG<math>\beta</math><sub>4</sub></b>		+	<b>ITG<math>\beta</math><sub>4</sub></b>		+
<b>ITG<math>\beta</math><sub>5</sub></b>		+	<b>ITG<math>\beta</math><sub>5</sub></b>		+
<b>ITG<math>\beta</math><sub>6</sub></b>		+	<b>ITG<math>\beta</math><sub>6</sub></b>		

Proteomic analysis of integrins in exosomes; '+' indicates positive for integrin expression by qualitative mass spectrometry.

# Gating machinery of InsP<sub>3</sub>R channels revealed by electron cryomicroscopy

Guizhen Fan<sup>1</sup>, Matthew L. Baker<sup>2</sup>, Zhao Wang<sup>2</sup>, Mariah R. Baker<sup>1</sup>, Pavel A. Sinyagovskiy<sup>1</sup>, Wah Chiu<sup>2</sup>, Steven J. Ludtke<sup>2</sup> & Irina I. Serysheva<sup>1</sup>

**Inositol-1,4,5-trisphosphate receptors (InsP<sub>3</sub>Rs) are ubiquitous ion channels responsible for cytosolic Ca<sup>2+</sup> signalling and essential for a broad array of cellular processes ranging from contraction to secretion, and from proliferation to cell death. Despite decades of research on InsP<sub>3</sub>Rs, a mechanistic understanding of their structure–function relationship is lacking. Here we present the first, to our knowledge, near-atomic (4.7 Å) resolution electron cryomicroscopy structure of the tetrameric mammalian type 1 InsP<sub>3</sub>R channel in its apo-state. At this resolution, we are able to trace unambiguously ~85% of the protein backbone, allowing us to identify the structural elements involved in gating and modulation of this 1.3-megadalton channel. Although the central Ca<sup>2+</sup>-conduction pathway is similar to other ion channels, including the closely related ryanodine receptor, the cytosolic carboxy termini are uniquely arranged in a left-handed  $\alpha$ -helical bundle, directly interacting with the amino-terminal domains of adjacent subunits. This configuration suggests a molecular mechanism for allosteric regulation of channel gating by intracellular signals.**

InsP<sub>3</sub>Rs belong to a superfamily of tetrameric cation channels that includes functionally distinct groups of ion channels (for example, K<sup>+</sup>, Na<sup>2+</sup>, Ca<sup>2+</sup>, transient receptor potential and cyclic nucleotide-gated channels). A common architectural feature of these channels is their central ion-permeation pore consisting of four membrane-spanning subunits or domains. However, these channels are quite different with respect to their activation and ability to respond to extracellular and intracellular stimuli, defining their particular roles in a wide range of cellular processes. Localized within the membranes of intracellular Ca<sup>2+</sup> stores such as the endoplasmic/sarcoplasmic reticulum, InsP<sub>3</sub>R channels have crucial roles in a variety of physiological functions, including gene transcription, fertilization, hormone secretion, metabolic regulation, immune responses, apoptosis, learning and memory. Their malfunction is associated with abnormal intracellular Ca<sup>2+</sup> levels linked to pathological conditions in humans, such as cardiac hypertrophy, heart failure, Alzheimer, Parkinson and Huntington diseases, cancer and stroke.

Type 1 InsP<sub>3</sub>R (InsP<sub>3</sub>R1) exemplifies the family of InsP<sub>3</sub>-gated Ca<sup>2+</sup> release channels, which is comprised of three homologous isoforms (types 1–3). InsP<sub>3</sub>R1 is the predominant Ca<sup>2+</sup> release channel in cerebellar Purkinje cells and best characterized member of the family. Despite its profound importance in physiology, a mechanistic basis for InsP<sub>3</sub>R1 function has remained elusive, largely owing to the lack of a detailed architecture of the intact InsP<sub>3</sub>R channel.

The quaternary structure of the entire InsP<sub>3</sub>R1 assembly was controversial<sup>1</sup> until our previous intermediate resolution electron cryomicroscopy (cryo-EM) structure<sup>2,3</sup>. While this structure is critical in providing the foundation for further structure–function studies, the resolution was insufficient to reveal the mechanistic features underlying the channel function. Furthermore, X-ray crystal structures for portions of InsP<sub>3</sub>R1 are available but limited to small soluble pieces (~15% of the 314-kilodalton protein)<sup>4–7</sup>. Knowledge of the full-length InsP<sub>3</sub>R architecture with atomistic details is paramount to understanding the molecular mechanism underlying channel function in both healthy and disease states. Here we present a cryo-EM structure

and model of the entire InsP<sub>3</sub>R1 channel in its ligand-free state determined at an overall 4.7 Å resolution by single-particle cryo-EM. The structure reveals distinct features that allow us to infer its gating mechanism.

## Overall architecture of InsP<sub>3</sub>R1

In this work, InsP<sub>3</sub>R1 from rat cerebellum was detergent-purified without adding InsP<sub>3</sub> or other channel-specific ligands, as previously described<sup>2</sup>. Before vitrification, InsP<sub>3</sub>R1 channel particles were stabilized by depletion of Ca<sup>2+</sup>, conditions that favour a closed state of the channel (Methods). The final map was reconstructed to 4.7 Å resolution (Extended Data Figs 1 and 2). The shape and overall dimensions of the tetrameric InsP<sub>3</sub>R1 3D density map are in excellent agreement with our previously published cryo-EM structure<sup>2</sup>: four subunits are arranged around a central channel axis forming two major regions, the bulky cytosolic region connected via ‘stalk’ densities to the transmembrane region (Extended Data Fig. 3 and Supplementary Video 1).

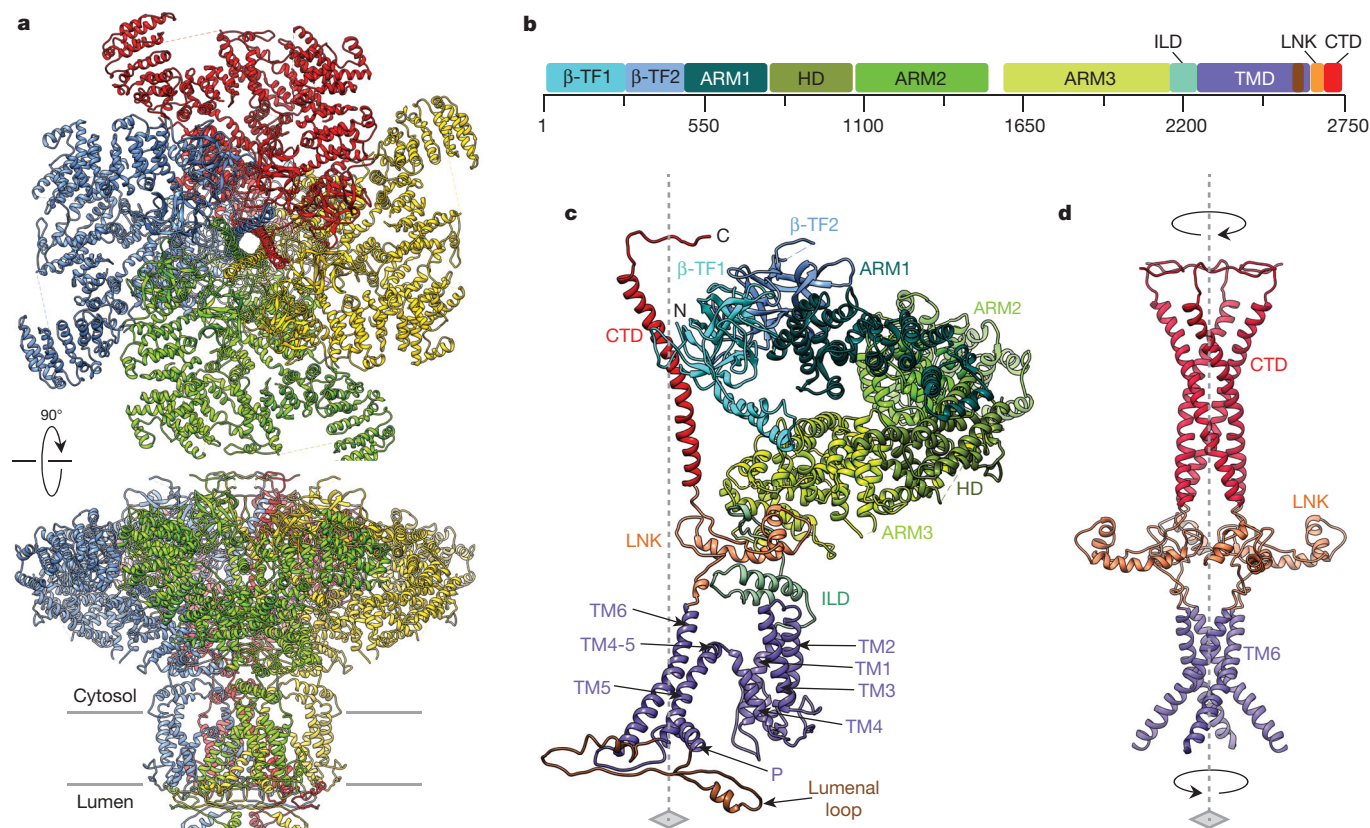
Resolution and resolvability of structural features are different in various parts of the map (Extended Data Fig. 2c) as suggested in our previous studies<sup>2,3</sup>. These differences are probably, at least in part, due to genuine flexibility associated with particular domains. Additionally, the potential presence of InsP<sub>3</sub>R1 splice variants may introduce some degree of compositional heterogeneity within the purified samples<sup>8–10</sup>. Thus, it is conceivable that these regions comprising alternatively spliced sequences represent some of the relatively poorly resolved regions of the map (Extended Data Figs 4a and 5). However, the transmembrane helices are well resolved in the map with many bulky side chains visible in the density map (Extended Data Fig. 6). Despite the observed variations in resolvability, we were able to model the backbone topology for 2,327 of 2,750 amino acids in the full-length InsP<sub>3</sub>R1 protein (Fig. 1).

## Subunit folds and tetrameric assembly

The individual InsP<sub>3</sub>R1 subunit is made up of ten domains (Fig. 1b, c), arranged within the tetrameric channel assembly around a central

<sup>1</sup>Department of Biochemistry and Molecular Biology, Structural Biology Imaging Center, The University of Texas Medical School at Houston, 6431 Fannin Street, Houston, Texas 77030, USA. <sup>2</sup>National Center for Macromolecular Imaging, Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA.





**Figure 1 | Overview of InsP<sub>3</sub>R1 structure.** **a**, Structure of InsP<sub>3</sub>R1 visualized in two orthogonal orientations: view from cytosol and side view along the membrane plane. Subunits are colour-coded. **b**, Linear representation of InsP<sub>3</sub>R1 structural domains (GI accession 17380349): domains are annotated

in the text, HD is the  $\alpha$ -helical domain. **c**, An individual subunit colour-coded by domain. **d**, Central core structure of tetrameric InsP<sub>3</sub>R1. Four-fold axis is indicated by the dashed line; arrows indicate the bundle handedness.

four-fold axis to fulfill specific functions, including channel stability, ion transport, ligand binding and regulation. In our model, ~90% of the protein sequence extends beyond the membrane into the cytosol including most of the N-terminal protein sequence and the C-terminal tail (Extended Data Fig. 5). The remaining portion encompasses the transmembrane and luminal domains.

At the amino terminus of each subunit, two contiguous  $\beta$ -trefoil domains ( $\beta$ -TF1 and  $\beta$ -TF2; residues 1–436) form apical densities around the central four-fold axis of the cytosolic region (Fig. 1b, c, Extended Data Fig. 4c and Supplementary Video 2). After  $\beta$ -TF2, there is a clear  $\alpha$ -helical pattern (residues 437–2192) forming three consecutive armadillo solenoid folds (ARM1–ARM3), with an  $\alpha$ -helical domain between ARM1 and ARM2. After ARM3, the subunit extends into the ‘intervening lateral’ domain (ILD, residues 2193–2272) that contains two anti-parallel  $\beta$ -strands followed by a helix–turn–helix motif. The C terminus of the ILD is connected to the transmembrane region, comprised of six  $\alpha$ -helices (TM1–TM6), a pore (P)-helix and three luminal loops. The C-terminal domain (CTD) contains an ~80 Å long  $\alpha$ -helix (residues 2681–2731) connected at its N terminus to TM6 via a helical linker domain (LNK, residues 2601–2680) (Fig. 1b–d).

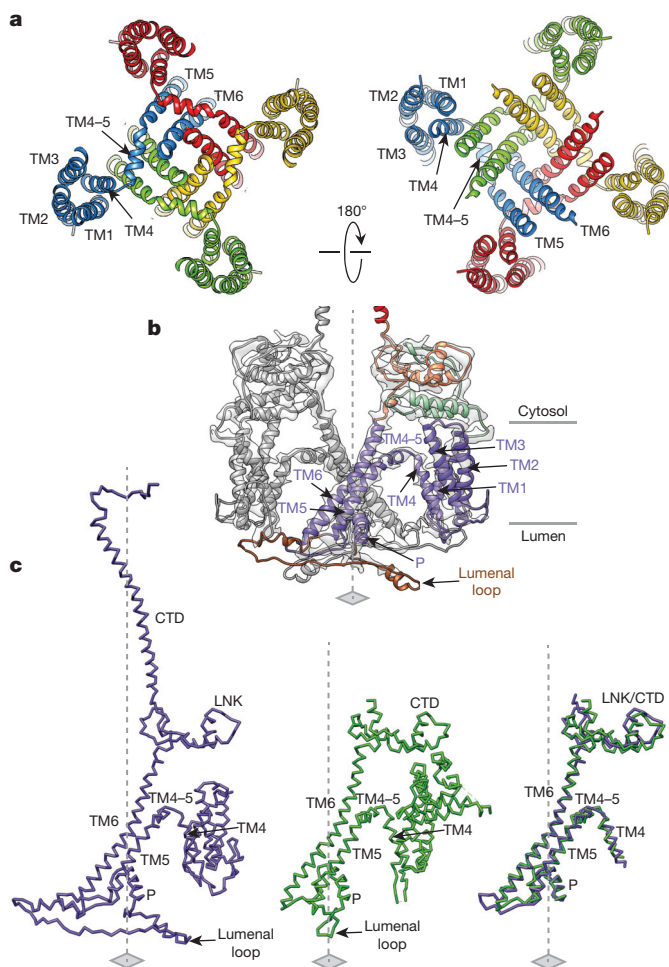
Unique to InsP<sub>3</sub>R1, the entire tetrameric architecture of the channel is built around two four-helix bundles: one transmembrane and one cytosolic bundle that together form a central core along the four-fold axis (Fig. 1d). The transmembrane bundle is right-handed and formed by the TM6 helices (one from each subunit), whereas the cytosolic bundle contains the CTD  $\alpha$ -helix from each subunit packed in a left-handed fashion. The latter spans most of the cytosolic region and earlier was described as the ‘plug’ density<sup>2</sup>. The transmembrane and cytosolic helical bundles are connected via four LNK domains

(one from each subunit), each consisting of two short, nearly orthogonal helices.

### Ca<sup>2+</sup> permeation pathway of InsP<sub>3</sub>R1

The cryo-EM densities comprising the transmembrane region of InsP<sub>3</sub>R1 are the most-resolved portions of the map (Extended Data Fig. 2c). As such, it was possible to construct a complete model for the transmembrane domains (TMDs), which constitute the ion-permeation pathway along the central four-fold axis (Fig. 2a, b). The Ca<sup>2+</sup> conduction path is lined by four TM6 helices, tilted ~37° relative to the membrane normal and packed in a right-handed bundle (Figs 2a, b and 3), a conformation originally observed in the KcsA channel<sup>11</sup>. The TM6 helices are ~55 Å long, extending beyond the cytosolic membrane surface and curving radially to form a tapering path for ions (Fig. 3a). The TM5 helices are packed against the TM6 helices (Fig. 2 and Supplementary Video 2). The luminal loop between the TM5 and TM6 helices contains a short P-helix (residues 2531–2544) and a selectivity filter comprising highly conserved residues 2546–2552 (Extended Data Figs 7a and 8c)<sup>2</sup>. The central bundle of TM5 and TM6 helices is linked to the flanking TM1–TM4 bundle via an amphipathic  $\alpha$ -helix (TM4–5), which lies parallel to the cytosolic membrane leaflet (Fig. 2). Mutations in the TM4–5 helix have been shown to affect the channel gating suggesting that this helix may influence the motions of the pore-lining helices of InsP<sub>3</sub>R (ref. 12). As observed in the other known tetrameric cation channels, TM4 cradles TM5 of the adjacent subunit (Fig. 2a).

The structure of the pore is wider at the luminal face of the membrane, where the P-helices and selectivity filter are located (Fig. 3). His2541 is present at the C-terminal end of the P-helix and forms a ring of positive charges on the luminal side, possibly repelling Ca<sup>2+</sup> in



**Figure 2 | Structure of the transmembrane domains.** **a**, Arrangement of transmembrane helices in tetrameric  $\text{InsP}_3\text{R1}$ ; subunits are colour-coded and viewed from cytosol (left) and lumen (right). **b**, The structures of two opposing TMDs are superimposed on cryo-EM densities; view is along the membrane plane. One subunit is colour-coded by domain. **c**, Structural comparison of TMDs from  $\text{InsP}_3\text{R1}$  (left) and  $\text{RyR1}$  (PDB accession 3J8H, middle); structurally conserved domains are shown overlapped (right); a root mean squared deviation (r.m.s.d.) value between the TMDs is 1.135 Å for 77 atom pairs.

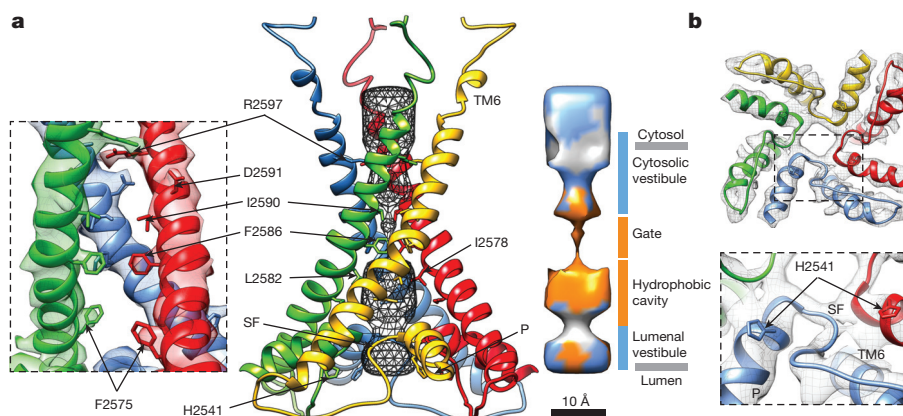
the non-conducting channel. In order for  $\text{Ca}^{2+}$  to pass beyond the luminal vestibule, we propose that the P-helices may undergo a structural rearrangement when the channel opens into a conductive state.

Our structure suggests that the physical gate for ion-permeation is located at the point of constriction along the TM6 helices (Fig. 3 and Extended Data Fig. 6c). This constriction, located closer to the cytosolic side of the membrane, includes a series of hydrophobic residues, Leu2582, Phe2586 and Ile2590, facing the  $\text{Ca}^{2+}$  permeation pathway. At Phe2586, densities for the side chains are clearly visible and point towards the central axis, where they shape a pore of  $\sim 5$  Å in diameter, suggesting a non-conducting channel conformation given an effective diameter of the hydrated  $\text{Ca}^{2+}$  (8–10 Å)<sup>13</sup>. Noteworthy, in this configuration the pore will also not be permissive for hydrated  $\text{K}^+$  ions (6 Å)<sup>14</sup>. Equivalent hydrophobic constrictions have been identified in the structures of other channels<sup>15–19</sup> (Supplementary Discussion). Just above the gate, a cytosolic vestibule formed by four TM6 helices has negatively charged residues that may facilitate  $\text{Ca}^{2+}$  translocation into the cytosol (Extended Data Fig. 7a). Altogether, our data support the idea of structural conservation of tetrameric cation channel design (Supplementary Discussion).

### Cytosolic scaffolding architecture

Perhaps the most prominent feature of the cytosolic region of  $\text{InsP}_3\text{R1}$  is the solenoid-like architecture of three  $\alpha$ -helical domains (ARM1–ARM3) formed by ensembles of armadillo repeats (Fig. 1 and Extended Data Fig. 4c, d). The modular architecture of the ARM domains in  $\text{InsP}_3\text{R1}$  is particularly amenable for generating different interfaces for recognition and binding of a large set of modulatory proteins whose putative binding sites have been previously identified<sup>20</sup> (Extended Data Fig. 4a). Thus, the flexible architecture of the ARM domains probably facilitates propagation of ligand-evoked signals towards the ion-conduction pathway.

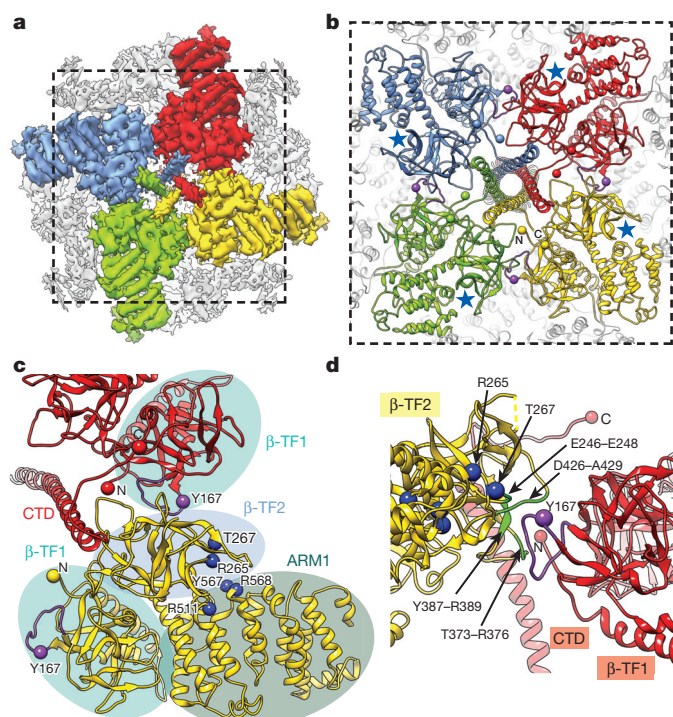
The ARM domains constitute peripheral densities of the cytosolic region that form a concave surface around the central helical bundle (Fig. 1c). ARM1 (residues 436–714) is formed by the stacking of five armadillo repeats (Extended Data Fig. 5). The first two armadillo repeats of the ARM1 and preceding two  $\beta$ -TF domains (Fig. 4a–c) constitute the ligand-binding domain (LBD)<sup>4–7</sup>. The  $\beta$ -TF1 domain (residues 5–225) was termed the  $\text{InsP}_3$ -binding suppressor domain, and  $\beta$ -TF2 (residues 226–435) with two  $\alpha$ -helices (residues 436–604) of ARM1 constitutes the  $\text{InsP}_3$ -binding core (IBC) region, also designated as  $\beta$ -IBC and  $\alpha$ -IBC subdomains, respectively (Fig. 4a–c and Extended Data Fig. 5). The suppressor domain,  $\beta$ -IBC and  $\alpha$ -IBC of each subunit form a triangular structure similar to previously reported crystal structures<sup>6,7</sup> (Extended Data Fig. 4b). The four LBDs constitute a cytoplasmic apical density around the CTD bundle (Fig. 4a). The ARM2 (residues 1030–1494) and ARM3 domains (residues 1594–2192) include ten and six armadillo repeats, respectively, and together with the helical domain (residues 715–1029), connect



**Figure 3 | Detailed structure of the  $\text{Ca}^{2+}$  conduction pathway.** **a**, A bundle of TM6 helices shape the ion permeation pathway. A series of hydrophobic residues within a constriction region of the channel pore are labelled. The wire cage is the void-density along the ion conduction pathway; colour-coded by

hydrophobicity (right). **b**, The luminal vestibule lined by P-helices and selectivity filter (SF) loops viewed along the four-fold axis from the lumen; insert shows the close-up view.





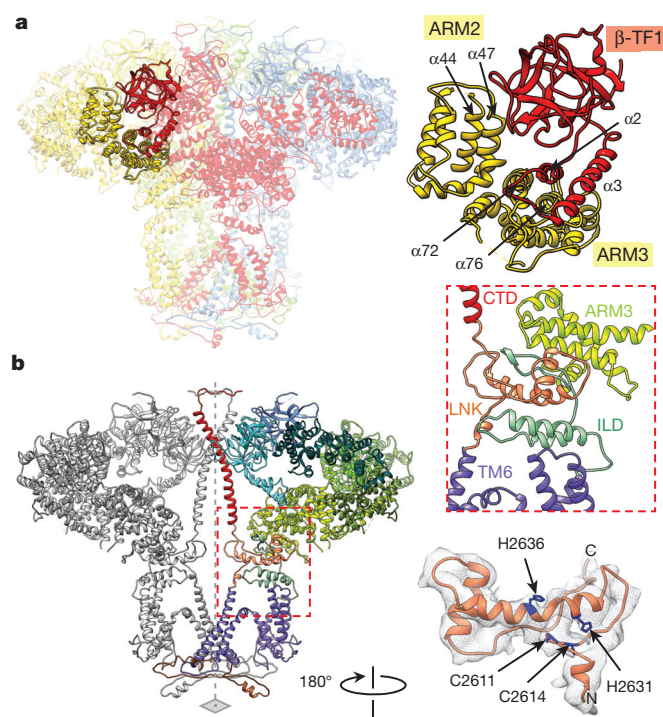
**Figure 4 | Structural coupling between the CTD and InsP<sub>3</sub>-binding domains.** **a**, Cryo-EM density map of InsP<sub>3</sub>R1 viewed from the cytosol with LBDs and CTDs colour-coded by subunit. **b**, InsP<sub>3</sub>R1 model corresponding to the region marked in **a**; blue stars denote InsP<sub>3</sub>-binding sites; purple spheres denote Tyr167. **c**, Contacts between LBD and CTD of adjacent subunits. Domains are denoted as ellipses colour-coded as in Fig. 1b. Residues involved in InsP<sub>3</sub> binding<sup>4–7</sup> are rendered as blue spheres. **d**, The interface between  $\beta$ -TF2 and  $\beta$ -TF1 of neighbouring subunits viewed orthogonally with respect to **c**. Residues on  $\beta$ -TF2 domain within 5 Å from hotspot loop of  $\beta$ -TF1 are coloured green.

the InsP<sub>3</sub>-binding domains to the channel-forming region. It is notable, that density in ARM2 is less resolved, suggesting higher structural flexibility (Extended Data Fig. 2c). The entire cytosolic solenoid scaffold in the tetrameric InsP<sub>3</sub>R1 communicates with the TMDs via the central helical bundle and the ILD (Fig. 1b, c and Supplementary Video 2).

### Gating via C- and N-terminal coupling

A functional hallmark of InsP<sub>3</sub>R channels is the complex interplay between the binding of primary ligands (InsP<sub>3</sub> and Ca<sup>2+</sup>) and channel gating. In addition, several intracellular regulatory molecules interact with InsP<sub>3</sub>R in a dynamic manner providing functional response in the channel<sup>20</sup> (Extended Data Fig. 4a). Our structure suggests that allosteric modulation of InsP<sub>3</sub>R gating by intracellular signals can occur via several different initiation points that mechanically couple these signals to a common ion-conduction pathway conserved among a large group of ion channels (Extended Data Fig. 7b).

Given the spatial separation between the InsP<sub>3</sub>-binding and ion-conduction domains within the tetrameric InsP<sub>3</sub>R1 (Fig. 1c and Extended Data Fig. 8a), there has to be a coupling mechanism for transmission of ligand-evoked signals to the pore to produce a specific gating event. In the crystal structures of the LBD, solved in both the InsP<sub>3</sub>-bound and apo-states<sup>6,7</sup>, InsP<sub>3</sub> binding at the cleft between the  $\beta$ -IBC and  $\alpha$ -IBC domains (Fig. 4b, c) causes closure of the ‘clam-like’ InsP<sub>3</sub>-binding pocket and movement of the suppressor domain. The intra-subunit interface between the  $\beta$ -TF1 and  $\beta$ -TF2 domains was shown to be dynamic, allowing for  $\beta$ -TF1 to twist in response to InsP<sub>3</sub> binding to the IBC<sup>6,7</sup>. While precise conformational changes associated with InsP<sub>3</sub>-induced activation of the channel gating in the native tetrameric InsP<sub>3</sub>R1 are not yet known, these signals may propagate

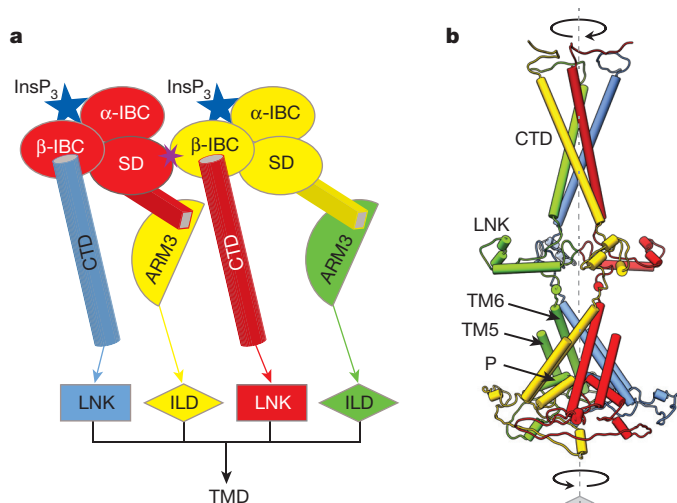


**Figure 5 | Cytosolic intra- and inter-subunit interactions.** **a**, Inter-subunit contacts of  $\beta$ -TF1 with ARM2 and ARM3 domains; subunits are colour-coded. **b**, Two opposing subunits are viewed along the membrane plane; domains of one subunit are colour-coded. Transmembrane–cytosolic domain interface is indicated with a red dashed line (left), and its close-up view is shown in insert (middle right). The structure of LNK domain is superimposed on cryo-EM densities (bottom right); residues of the C2H2 Zn<sup>2+</sup>-finger-like motif are coloured blue.

towards the TMDs via discrete intra- and inter-subunit interfaces identified in our structure (Figs 4 and 5). Our structure shows that the helix–turn–helix ( $\alpha$ 2–turn– $\alpha$ 3) motif, probably communicates with the  $\alpha$ 72 and  $\alpha$ 76 helices of the ARM3 domain within the adjacent subunit (Fig. 5a and Extended Data Fig. 5). Because the C terminus of the ARM3 domain is connected to the TMD via the ILD, it is conceivable that interactions between the  $\beta$ -TF1 and ARM3 domains will propagate InsP<sub>3</sub> evoked conformational changes in the LBD to the pore (Fig. 6a). Importantly, suppressor domain deletion mutants of InsP<sub>3</sub>R1 have been shown not to exhibit any measurable InsP<sub>3</sub>-induced Ca<sup>2+</sup> release<sup>21</sup>, emphasizing its critical role in transmitting the InsP<sub>3</sub> binding signal. Furthermore, the ARM3 domain contains a putative Ca<sup>2+</sup>-sensor region (residues 1933–2271), where a highly conserved Glu2101 has been implicated in mediating channel gating<sup>22,23</sup>. However, no EF-hand Ca<sup>2+</sup> binding motif has been found in InsP<sub>3</sub>R1 (Extended Data Fig. 9).

The  $\beta$ -TF1 domain also forms inter-subunit interactions with  $\beta$ -TF2 and ARM2 domains. The loop of  $\beta$ -TF1 (residues 166–180) is positioned to interact with non-contiguous regions of  $\beta$ -TF2 of the neighbouring subunit including residues 246–248, 373–376, 387–389 and 426–429 (Fig. 4d). Mutation of Tyr167 was shown to impair InsP<sub>3</sub>-induced Ca<sup>2+</sup> release but not affect InsP<sub>3</sub> binding to InsP<sub>3</sub>R1 (ref. 24), suggesting the importance of this inter-subunit interface in signal transmission. Furthermore, the  $\beta$ -TF1 loop (residues 136–139) makes additional inter-subunit contacts with  $\alpha$ 44 and  $\alpha$ 47 helices of the ARM2 domain (Fig. 5a), which specifies modulatory activities for InsP<sub>3</sub>-mediated Ca<sup>2+</sup> release<sup>25</sup> (Extended Data Fig. 4a). All together, this alludes to the importance of the  $\beta$ -TF1 domain in coupling InsP<sub>3</sub> binding to activation of the channel gate. However, our structure shows no direct coupling between the LBD and TM4–5 as previously proposed<sup>26</sup>, suggesting a more long-range mechanism for signal transmission.





**Figure 6 | Model for coupling mechanism of InsP<sub>3</sub>R1 activation by InsP<sub>3</sub>.** **a**, Schematic representation of inter-subunit contacts involved in propagation of InsP<sub>3</sub>-binding signal to the pore (colour-coded by subunit): from  $\beta$ -IBC to CTD/LNK domains and from the suppressor domain (SD) to ARM3/ILD domains of neighbouring subunits. **b**, The InsP<sub>3</sub>-induced changes in the LBDs can cause the helices in the cytosolic bundle to rearrange and trigger the motions of the LNKs, that may force the transmembrane bundle to adopt a conformation permeable for Ca<sup>2+</sup>; colour-coded by subunit.

The CTD in InsP<sub>3</sub>R1 is comprised of a coiled-coil motif, forming a long helical bundle that spans almost the entire cytosolic assembly of the channel, unlike the CTD in ryanodine receptor 1 (RyR1)<sup>27–29</sup> (Figs 1c, d, 2c and Extended Data Fig. 8a). The coiled-coil structure is potentially stabilized by a predicted saltbridge between adjacent subunits at residues Gln2700 and Lys2701 (Extended Data Fig. 10a). Furthermore, the CTD of one subunit interacts with the  $\beta$ -TF2 of the adjacent subunit by predominantly inter-domain electrostatic interactions (Fig. 4 and Extended Data Fig. 10b, c). This observation is consistent with earlier biochemical studies of InsP<sub>3</sub>R proposing that the N-terminal and C-terminal regions are probably in a close association in the native channel<sup>30,31</sup>. In this arrangement, the cytosolic helical bundle is in a prime position to sense a signal resulting from InsP<sub>3</sub> binding and may undergo a conformational change playing a critical role in transmitting the InsP<sub>3</sub> signal to the TMDs (Figs 4 and 6).

The CTDs are connected to LNK domains that form a connecting ring between the cytosolic and transmembrane bundles and essentially establish continuous communication between the N-terminal LBDs and the pore-lining TM6 helices (Figs 1c, d and 5b). Noteworthy, the LNK domain of InsP<sub>3</sub>R1 is structurally analogous to the CTD domain (residues 4957–5037) in RyR1 (refs 21, 32). Both domains contain a Cys2His2 (C2H2) Zn<sup>2+</sup>-finger motif, although in InsP<sub>3</sub>R1 there are 16 residues between Cys2614 and His2631 rather than the typical 12 residues (Fig. 5b and Extended Data Fig. 7a). While modulation of the InsP<sub>3</sub>R channel function has not been investigated regarding its sensitivity to Zn<sup>2+</sup>, this domain may have a role in coordinating metals that influence channel gating. Mutational studies of the C2H2 residues were shown to either inhibit (Cys2611Ser, Cys2614Ser, His2636Ala) or completely abolish (His2631Ala) channel function, indicating that the LNK domain is positioned to be a crucial structural component for allosteric modulation of channel gating<sup>21,32</sup>.

This region also contains the sequence (residues 2630–2655) proposed to be essential for InsP<sub>3</sub>R tetramerization<sup>33</sup> (Extended Data Fig. 7a). A predicted hydrogen bond between Asn2241 and Glu2617 in our structure may bridge the LNK and ILD of neighbouring subunits, possibly having a modest role in channel formation. However, the larger role of the LNK is not entirely evident from our structure given the lack of direct inter-subunit contacts of the LNK domains in the tetrameric InsP<sub>3</sub>R1 assembly. Rather, the combined interactions

throughout the entire central core structure of InsP<sub>3</sub>R1 may serve to stabilize the tetrameric interfaces providing structural and functional integrity of InsP<sub>3</sub>R channel (Fig. 6b).

The dominant role of the CTD in transmitting the conformational changes triggered by InsP<sub>3</sub> binding to channel gate opening is further supported by an earlier study demonstrating that deletion of 43 residues from the CTD disrupts channel gating<sup>26</sup>. Furthermore, in a chimaeric channel, comprising the TMD of RyR1 and the cytosolic domains of InsP<sub>3</sub>R1, the InsP<sub>3</sub> efficacy is substantially decreased while InsP<sub>3</sub> binding is minimally affected<sup>7</sup>. It is conceivable that the lack of the InsP<sub>3</sub>R1 CTD bundle in the chimaeric channel causes uncoupling of InsP<sub>3</sub>-evoked conformational changes to the channel gating. The unique CTD architecture of InsP<sub>3</sub>R1 suggests that the two families of Ca<sup>2+</sup> release channels explore different mechanisms for transmitting ligand-evoked signals to the ion-permeation pore (Supplementary Discussion). Our hypothetical model for transmission of the InsP<sub>3</sub>-evoked signal is schematically shown in Fig. 6. Again, while the precise InsP<sub>3</sub>-induced conformational changes remain to be revealed, the channel activation relies on concerted inter- and intra-domain interactions mediated in the context of the tetrameric InsP<sub>3</sub>R1 assembly. In this scenario, binding of a single InsP<sub>3</sub> molecule to one subunit can trigger a cascade of conformational changes in domains of two neighbouring subunits to propagate the activating signal to the ion-conducting pore.

## Conclusion

Here we present the first, to our knowledge, near-atomic resolution structure of a eukaryotic intracellular InsP<sub>3</sub>-gated Ca<sup>2+</sup> release channel purified from rat cerebellum. Our structural analysis reveals conservation of the ion-conduction pore across the tetrameric cation channel family, further suggesting a conserved mechanism for translocation of ions in these channels. However, the control of channel gating in a cellular context is conferred by additional domains that are attached to the pore and contain specific functions, such as voltage-sensing or ligand-binding. The distinctive molecular architecture of these domains defines the nature of work that needs to be performed to convert cellular signals into mechanical gating motion. Thus, while InsP<sub>3</sub>R1 shares a set of common features with the closely related RyR1 channel, the unique architecture in its C-terminal domain suggests a distinctive allosteric mechanism underlying activation of InsP<sub>3</sub>R gating (Fig. 6). This mechanism requires direct mechanical coupling between the C terminus and InsP<sub>3</sub>-binding domains of adjacent subunits and involves rearrangements of the inter-domain interfaces identified in this study. To our knowledge, this is the first structural evidence for the recruitment of the C terminus in gating of the InsP<sub>3</sub>R channel.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 23 February; accepted 24 July 2015.**

**Published online 12 October 2015.**

- Serysheva, I. I. & Ludtke, S. J. 3D structure of IP<sub>3</sub> receptor. *Curr. Topics Memb.* **66C**, 171–189 (2010).
- Ludtke, S. J. *et al.* Flexible architecture of IP<sub>3</sub>R1 by Cryo-EM. *Structure* **19**, 1192–1199 (2011).
- Murray, S. C. *et al.* Validation of cryo-EM structure of IP<sub>3</sub>R1 channel. *Structure* **21**, 900–909 (2013).
- Bosanac, I. *et al.* Structure of the inositol 1,4,5-trisphosphate receptor binding core in complex with its ligand. *Nature* **420**, 696–700 (2002).
- Bosanac, I. *et al.* Crystal structure of the ligand binding suppressor domain of type 1 inositol 1,4,5-trisphosphate receptor. *Mol. Cell* **17**, 193–203 (2005).
- Lin, C. C., Baek, K. & Lu, Z. Apo and InsP-bound crystal structures of the ligand-binding domain of an InsP receptor. *Nature Struct. Mol. Biol.* **18**, 1172–1174 (2011).
- Seo, M. D. *et al.* Structural and functional conservation of key domains in InsP<sub>3</sub> and ryanodine receptors. *Nature* **483**, 108–112 (2012).
- Nakagawa, T., Okano, H., Furuichi, T., Aruga, J. & Mikoshiba, K. The subtypes of the mouse inositol 1,4,5-trisphosphate receptor are expressed in a tissue-specific and

- developmentally specific manner. *Proc. Natl Acad. Sci. USA* **88**, 6244–6248 (1991).
9. Nucifora, F. C. Jr, Li, S. H., Danoff, S., Ullrich, A. & Ross, C. A. Molecular cloning of a cDNA for the human inositol 1,4,5-trisphosphate receptor type 1, and the identification of a third alternatively spliced variant. *Brain Res. Mol. Brain Res.* **32**, 291–296 (1995).
  10. Danoff, S. K. *et al.* Inositol 1,4,5-trisphosphate receptors: distinct neuronal and nonneuronal forms derived by alternative splicing differ in phosphorylation. *Proc. Natl Acad. Sci. USA* **88**, 2951–2955 (1991).
  11. Doyle, D. A. *et al.* The structure of the potassium channel: molecular basis of K<sup>+</sup> conduction and selectivity. *Science* **280**, 69–77 (1998).
  12. Schug, Z. T. *et al.* Molecular characterization of the inositol 1,4,5-trisphosphate receptor pore-forming segment. *J. Biol. Chem.* **283**, 2939–2948 (2008).
  13. Fulton, J., Heald, S., Baday, Y. & Simonson, J. Understanding the effects of concentration on the solvation structure of Ca<sup>2+</sup> in aqueous solution. I: the perspective on local structure from EXAFS and XANES. *J. Phys. Chem.* **107**, 4688–4696 (2003).
  14. Gillespie, D., Xu, L. & Meissner, G. Selecting ions by size in a calcium channel: the ryanodine receptor case study. *Biophys. J.* **107**, 2263–2273 (2014).
  15. Kuo, A. *et al.* Crystal structure of the potassium channel KirBac1.1 in the closed state. *Science* **300**, 1922–1926 (2003).
  16. Long, S. B., Campbell, E. B. & Mackinnon, R. Crystal structure of a mammalian voltage-dependent Shaker family K<sup>+</sup> channel. *Science* **309**, 897–903 (2005).
  17. Tao, X., Avalos, J. L., Chen, J. & MacKinnon, R. Crystal structure of the eukaryotic strong inward-rectifier K<sup>+</sup> channel Kir2.2 at 3.1 Å resolution. *Science* **326**, 1668–1674 (2009).
  18. Miyazawa, A., Fujiyoshi, Y. & Unwin, N. Structure and gating mechanism of the acetylcholine receptor pore. *Nature* **423**, 949–955 (2003).
  19. Chang, G., Spencer, R. H., Lee, A. T., Barclay, M. T. & Rees, D. C. Structure of the MscL homolog from *Mycobacterium tuberculosis*: a gated mechanosensitive ion channel. *Science* **282**, 2220–2226 (1998).
  20. Serysheva, I. I. Toward a high-resolution structure of IP<sub>3</sub>R channel. *Cell Calcium* **56**, 125–132 (2014).
  21. Uchida, K., Miyauchi, H., Furuichi, T., Michikawa, T. & Mikoshiba, K. Critical regions for activation gating of the inositol 1,4,5-trisphosphate receptor. *J. Biol. Chem.* **278**, 16551–16560 (2003).
  22. Tu, H. *et al.* Functional and biochemical analysis of the type 1 inositol (1,4,5)-trisphosphate receptor calcium sensor. *Biophys. J.* **85**, 290–299 (2003).
  23. Miyakawa, T. *et al.* Ca<sup>2+</sup>-sensor region of IP<sub>3</sub> receptor controls intracellular Ca<sup>2+</sup> signaling. *EMBO J.* **20**, 1674–1680 (2001).
  24. Yamazaki, H., Chan, J., Ikura, M., Michikawa, T. & Mikoshiba, K. Tyr-167/Trp-168 in type 1/3 inositol 1,4,5-trisphosphate receptor mediates functional coupling between ligand binding and channel opening. *J. Biol. Chem.* **285**, 36081–36091 (2010).
  25. Soulsby, M. D., Alzayady, K., Xu, Q. & Wojcikiewicz, R. J. The contribution of serine residues 1588 and 1755 to phosphorylation of the type I inositol 1,4,5-trisphosphate receptor by PKA and PKG. *FEBS Lett.* **557**, 181–184 (2004).
  26. Schug, Z. T. & Joseph, S. K. The role of the S4–S5 linker and C-terminal tail in inositol 1,4,5-trisphosphate receptor function. *J. Biol. Chem.* **281**, 24431–24440 (2006).
  27. Zalk, R. *et al.* Structure of a mammalian ryanodine receptor. *Nature* **517**, 44–49 (2015).
  28. Yan, Z. *et al.* Structure of the rabbit ryanodine receptor RyR1 at near-atomic resolution. *Nature* **517**, 50–55 (2015).
  29. Efremov, R. G., Leitner, A., Aebersold, R. & Raunser, S. Architecture and conformational switch mechanism of the ryanodine receptor. *Nature* **517**, 39–43 (2015).
  30. Boehning, D. & Joseph, S. K. Direct association of ligand-binding and pore domains in homo- and heterotetrameric inositol 1,4,5-trisphosphate receptors. *EMBO J.* **19**, 5450–5459 (2000).
  31. Yoshikawa, F., Iwasaki, H., Michikawa, T., Furuichi, T. & Mikoshiba, K. Trypsinized cerebellar inositol 1,4,5-trisphosphate receptor. Structural and functional coupling of cleaved ligand binding and channel domains. *J. Biol. Chem.* **274**, 316–327 (1999).
  32. Bhanumathy, C., da Fonseca, P. C., Morris, E. P. & Joseph, S. K. Identification of functionally critical residues in the channel domain of inositol trisphosphate receptors. *J. Biol. Chem.* **287**, 43674–43684 (2012).
  33. Galvan, D. L. & Mignery, G. A. Carboxyl-terminal sequences critical for inositol 1,4,5-trisphosphate receptor subunit assembly. *J. Biol. Chem.* **277**, 48248–48260 (2002).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank S. Murray for his assistance with RELION1.3. This work was supported by grants from the National Institutes of Health (R01GM072804, R21AR063255, S10OD016279, P41GM103832, R01GM079429, R01GM080139 and R21GM100229), the American Heart Association (14RNT1980029), the Muscular Dystrophy Association (295138) and National Science Foundation (DBI-1356306). We gratefully acknowledge the assistance and computing resources provided by the Center for Computational and Integrative Biomedical Research of Baylor College of Medicine and the Texas Advanced Computing Center at the University of Texas at Austin in the completion of this work.

**Author Contributions** I.I.S. conceived the project; G.F. and I.I.S. purified and characterized the protein; G.F., Z.W. and I.I.S. performed cryo-EM experiments, including cryospecimen preparation and data acquisition; G.F., Z.W., P.A.S., S.J.L. and I.I.S. analysed data; M.L.B. and M.R.B. built an atomic model; M.L.B., M.R.B., I.I.S. and W.C. interpreted the model; I.I.S., W.C., M.L.B. and M.R.B. wrote a manuscript with contributions from all authors.

**Author Information** Cryo-EM density map of InsP<sub>3</sub>R1 has been deposited in the Electron Microscopy Data Bank under accession number EMD-6369. The model coordinates have been deposited in the Protein Data Bank under accession number 3JAV. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to I.I.S. ([irina.i.serysheva@uth.tmc.edu](mailto:irina.i.serysheva@uth.tmc.edu)).

## METHODS

Statistical methods for predetermining sample sizes are not appropriate for cryo-EM analysis, which typically involves  $10^4$ – $10^6$  images. Instead, a range of more sophisticated validations and error estimates were performed as described below.

**Cryo-EM data acquisition.** Detergent-solubilized InsP<sub>3</sub>R1 was purified from rat cerebellum, and its structure–function integrity has been confirmed as described in our earlier study<sup>2</sup>. The absence of any channel-specific modulatory proteins has been confirmed using immunoprecipitation/mass spectrometry (data not shown)<sup>34</sup>. The vitrification of the purified protein was performed as previously described<sup>2</sup>. Images of frozen-hydrated InsP<sub>3</sub>R1 particles were acquired on a Technai G2 Polara electron microscope (FEI) operated at 300 kV using a K2 Summit direct electron detector camera (Gatan). Images were collected in dose fractionation super-resolution counting mode at a nominal magnification of  $\times 23,000$ , corresponding to a calibrated physical pixel size of 1.62 Å and super-resolution pixel size of 0.81 Å. The dose rate on the camera was set to  $\sim 10$  electrons pixel<sup>-1</sup> s<sup>-1</sup>. The total exposure time was 6 s, leading to a total accumulated dose of 22 electrons Å<sup>-2</sup> on the specimen. Each image stack was fractionated into 30 subframes, each with an accumulation time of 0.2 s per frame. Images were acquired at the defocus range of  $-0.6$  to  $-3.5$  µm.

**Image processing and 3D reconstruction.** Dose-fractionated super-resolution raw image stacks of ice-embedded InsP<sub>3</sub>R1 were binned  $2 \times 2$  by Fourier cropping resulting in a pixel size of 1.62 Å for further image processing. Each image stack was subjected to motion correction using ‘dosefgpu\_driftcorr’<sup>35</sup>, and a sum of subframes 1–29 in each image stack was used for further image processing (Extended Data Fig. 1). We used ‘e2evalimage.py’ in EMAN 2.1 package to select 3,743 micrographs from a total of 4,160 micrographs for subsequent processing. The signal in the motion-corrected images extends beyond  $\sim 4$  Å (Extended Data Fig. 1b).

156,805 particles were boxed out manually using ‘e2boxer.py’. Image processing was then performed independently in both RELION1.3 and EMAN 2.1 beginning with the same set of boxed out particles. For the RELION1.3 reconstruction, defocus and astigmatism were determined for each micrograph by CTFFIND3 (ref. 36). Our previously published map (EMDB accession 5278) was low pass filtered to 60 Å resolution and used as a starting model for the RELION1.3 refinement. The first refinement yielded subnanometer resolution. After this step, we ran several rounds of iterative 3D classification and 3D auto-refinement to extract the most self-consistent subset of the particle data set. The final map was generated from 96,106 particles.

The EMAN2.1 reconstruction followed standard refinement procedures<sup>2,37</sup> with an initial model generated from the current data set with no reference to the prior published structure. On the basis of per-particle SSNR estimates, the best 105,000 particles were included as input to the refinement. During refinement, the worst  $\sim 30\%$  of these particles were discarded from each reference-based class-average, based on mutual similarity among particles. Refinement used the EMAN2.1 ‘e2refine\_easy’ script, in which most refinement parameters were automatically selected. An initial low resolution refinement was performed using a smaller subset of the data to improve the initial model. As a result of image quality improvement, 2D class-average images generated through iterative image processing exhibit significantly more details than in previous studies<sup>2</sup> (Extended Data Fig. 1c).

The particle orientation distribution is not entirely uniform with a broad distribution focused on a  $\sim 45^\circ$  wide distribution near the equator (side views) and another grouping near the top view, along the axis of symmetry (Extended Data Fig. 2d). Given the C4 group symmetry, this distribution includes considerable numbers of particles extending from the equator to the pole, which is sufficient for a complete reconstruction with no missing information.

The final RELION1.3 and EMAN2.1 refinements both used gold standard procedures<sup>38,39</sup>. In this method, the particle data were split into even/odd numbered halves, and two maps were refined completely independently. The Fourier shell correlation (FSC) 0.143 cut-off was used to estimate the resolution of final 3D reconstructions with a soft auto-mask in RELION post-processing (Extended Data Fig. 2a). RELION1.3 density map was sharpened by applying a B-factor of  $-256$  Å<sup>2</sup> that was estimated using an automated procedure<sup>40</sup> and visualized with Chimera<sup>41</sup>. EMAN2.1 maps were automatically filtered as part of the refinement. Local resolution variations were estimated using ResMap<sup>42</sup> (Extended Data Fig. 2c).

The resulting 3D reconstructions from the two software packages, RELION1.3 and EMAN2.1, were assessed by computing an FSC between these two maps. Since both maps used the full data set, it is more appropriate to use a 0.25 FSC threshold in this comparison, which corresponds to  $2\times$  the SSNR of the 0.143 threshold (Extended Data Fig. 2b). Note, that regardless of localized differences at the highest resolutions, the topology of the two maps is consistent, and the same

protein model can be optimized for both maps. All structure/function interpretations would be identical for both maps, irrespective of the resolution value ascribed to the maps.

**Model building.** Before building a model for the InsP<sub>3</sub>R1 subunit, secondary structure identification was performed with SSEHunter in Gorgon<sup>43,44</sup>. A total of 106 helices were identified and a series of  $\beta$ -sheets in the central apical domains of cytosolic region were identified per InsP<sub>3</sub>R1 subunit. Sequence analysis was then performed on the InsP<sub>3</sub>R1 primary sequence (GI accession 17380349). As the entire protein was generally too large for most web-based services, serial overlapping sequence segments were used to initially screen the sequence; sequence segments were  $\sim 800$  amino acids in length and overlapped by  $\sim 200$  amino acids in either direction. Secondary structure prediction was performed using JPRED3<sup>45</sup>. Analysis of the secondary structure illustrated a predominance of helices, except over the first  $\sim 500$  amino acids which were indicated to be mainly  $\beta$ -sheet. Deriving the molecular structure of InsP<sub>3</sub>R1 used a three-pronged approach to generate initial models for different parts of the protein sequence depending on the availability of available crystal structures, fold recognition using homologous structures and completely *de novo* modelling (Extended Data Fig. 4a). Fold recognition was first performed on the InsP<sub>3</sub>R1 sequence using PHYRE2 (ref. 46). As expected, the known crystal structures of the N-terminal domains of InsP<sub>3</sub>R (PDB accessions 3T8S and 3UJ4) were identified spanning residues 7–580. In addition, the structure of the RyR1 N-terminal domain was identified (PDB accession 2XOA). Two structural homologues over residues 1104–1431 were identified; the ARM repeat from 1N4K and an  $\alpha$ - $\alpha$  superhelix/ARM repeat (PDB accession 1XQR) with 99.7% (25% identity) and 42% (12% sequence identity) confidence, respectively. Models for these domains were obtained directly from the PHYRE2 website (intensive option) and fit to the InsP<sub>3</sub>R1 density map using UCSF Chimera’s ‘fit in map’ module, EMAN’s Foldhunter and Gorgon<sup>44,47</sup>. The N-terminal model was localized to LBD and ARM1 regions of the density map. The composite model for residues 1104–1431, containing an  $\alpha$ - $\alpha$  superhelix motif, was localized in the ARM2 region of the density map. The secondary structure elements in the models matched well with those identified in the cryo-EM density map, further validating the fit of these structures. Together, these two models accounted for  $\sim 33\%$  of the total amino acids in the InsP<sub>3</sub>R1 subunit. A third homologous domain, the crystal structure of a voltage-gated Na<sup>+</sup> channel (PDB accession: 4DXW), was found using PHYRE corresponding to the TMD residues (residues 2278–2605) with 98.7% confidence and  $\sim 19\%$  sequence identity. While the structural homologue was fit to the TMD of our InsP<sub>3</sub>R1 cryo-EM density map, it was not directly used for modelling. The voltage-gated Na<sup>+</sup> channel structure was used to confirm the putative InsP<sub>3</sub>R1 subunit boundaries in the TMD. Modelling of the TMD and the remaining residues were accomplished *de novo* as described below.

The remainder of the sequence for InsP<sub>3</sub>R1 was modelled using our previously developed *de novo* modelling protocol<sup>48,49</sup> (Extended Data Fig. 4a). Essentially, we used the aforementioned structural homologues to anchor our sequence assignments. Extending from the ends of the models, we identified the next directly connected secondary structure element via density and/or density skeleton. We then attempted to match the observed SSEHunter secondary structure elements to what was predicted in the primary sequence; the fitted models provided a rough registration of the sequence in the density map, making this process feasible. A C $\alpha$  backbone model was then constructed for this region using either Gorgon<sup>50</sup> or Coot<sup>51</sup>. The modelling procedure was repeated extending from the previously placed secondary structure elements until either the models merged or encountered a break in the density. It should be noted that model construction was done with the aforementioned 4.7 Å resolution RELION1.3 map, preserving the EMAN2.1 map to act as a control. The C $\alpha$  backbone model was subsequently verified using our Pathwalking protocol<sup>52</sup>, which helps to identify alternate backbone topologies. Through this process, less than 15% of the residues were not modelled, likely due to the poor resolvability of the map in those regions (Extended Data Fig. 4a; see main text).

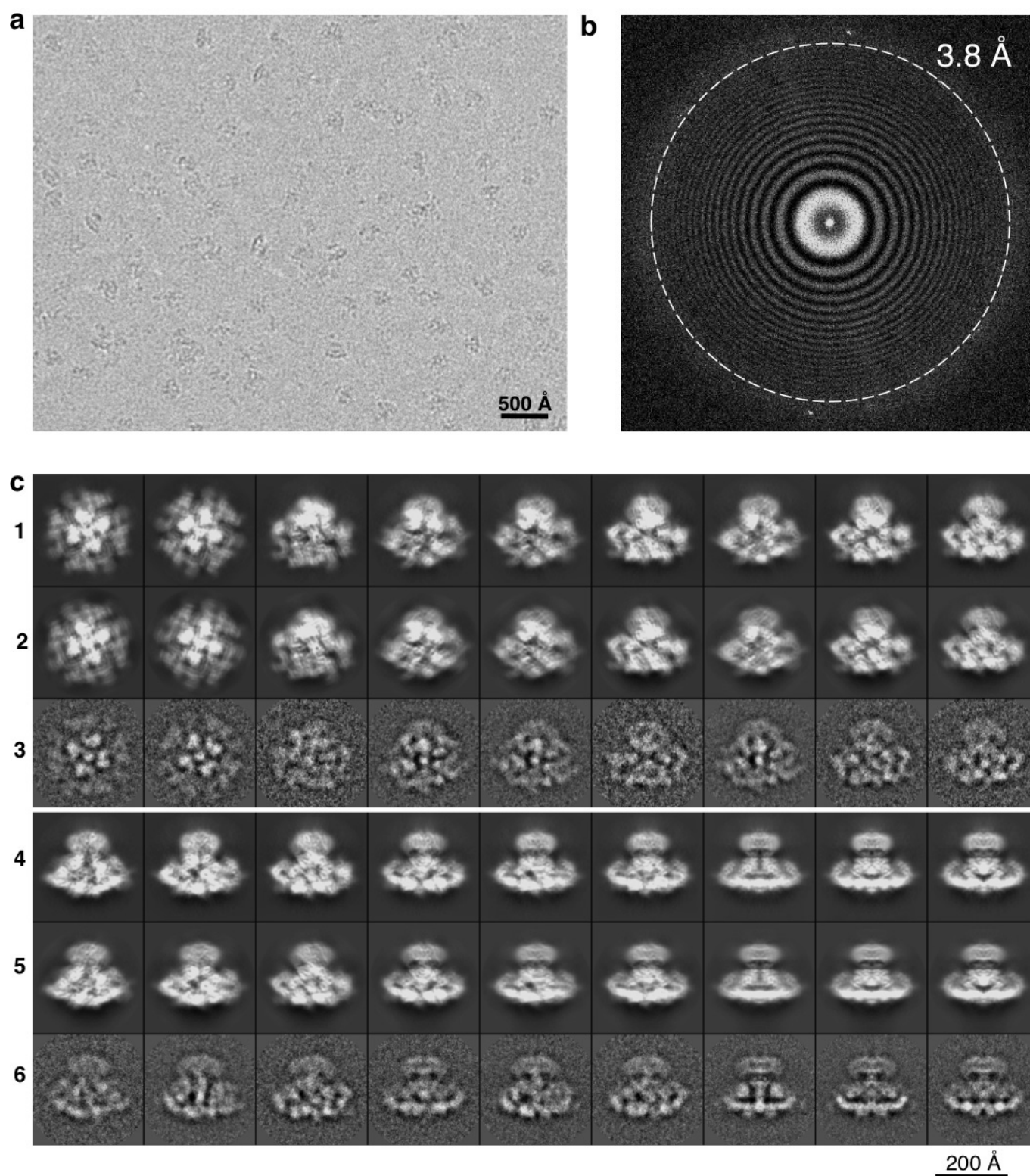
Once a complete C $\alpha$  backbone model was constructed, the model was transformed into a full atom model using REMO<sup>53</sup> and manually adjusted to best fit the density in Coot<sup>51</sup>. As the density in the transmembrane region was best resolved, model optimization to the density began in the transmembrane regions followed by the remaining regions. From this initial Coot-optimized model, the full-atom model for the tetrameric assembly was then refined with Phenix’s real-space refinement tools as previously described<sup>54</sup>. In brief, we first used the default real-space refinement settings, followed by subsequent rounds of refinement ( $\sim 10$ ) incorporating simulated annealing, global minimization, rigid body minimization and local grid searching. After Phenix refinement, the model was again manually adjusted in Coot to best fit the density. This process was iterated over 10 rounds of map-model refinement with Phenix and Coot. As a final test of model fidelity and map agreement, the final model refined against the



RELION1.3 map was then in turn refined against the EMAN2.1 map. This resulting refinement gave a nearly identical model with an r.m.s.d. of only 0.8 Å. Inter- and intra-subunit interfaces described here were identified using PDBSum<sup>55</sup> and PDBe PISA<sup>56</sup>.

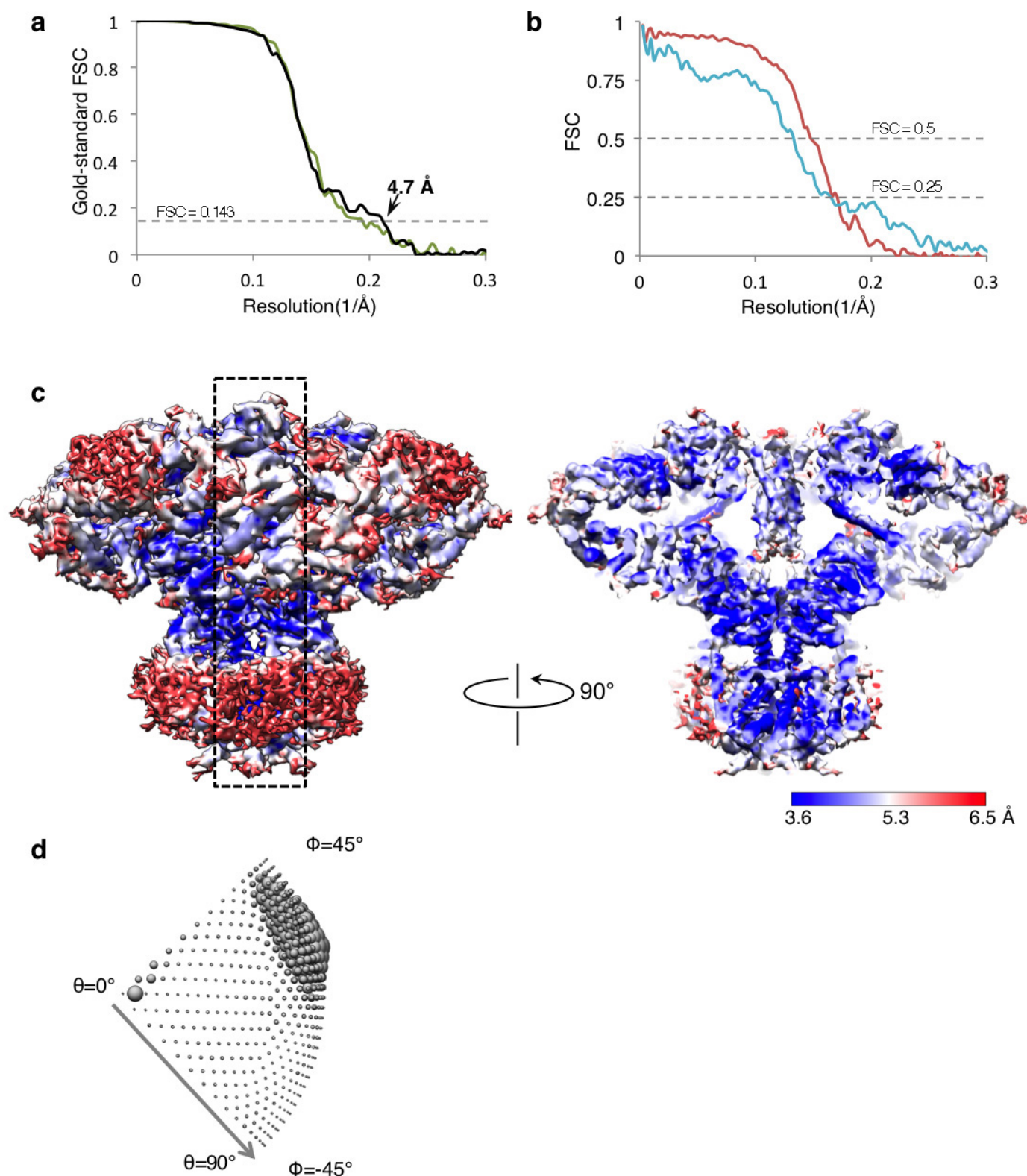
It should be noted that the *de novo* modelling process relies on the ability to register sequence predicted helices with those found in the structure. This registration is further enforced by anchoring large, bulky side chains to protruding densities in helices. However, potential errors in the map, modelling and refinement procedures may result in register shifts, non-optimal side-chain placement and tracing errors where there are not sufficient density anchors or where the map is not as well resolved<sup>50</sup>.

34. Malovannaya, A. *et al.* Streamlined analysis schema for high-throughput identification of endogenous protein complexes. *Proc. Natl Acad. Sci. USA* **107**, 2431–2436 (2010).
35. Li, X. *et al.* Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nature Methods* **10**, 584–590 (2013).
36. Mindell, J. A. & Grigorieff, N. Accurate determination of local defocus and specimen tilt in electron microscopy. *J. Struct. Biol.* **142**, 334–347 (2003).
37. Tang, G. *et al.* EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol.* **157**, 38–46 (2007).
38. Henderson, R. *et al.* Outcome of the first electron microscopy validation task force meeting. *Structure* **20**, 205–214 (2012).
39. Scheres, S. H. & Chen, S. Prevention of overfitting in cryo-EM structure determination. *Nature Methods* **9**, 853–854 (2012).
40. Rosenthal, P. B. & Henderson, R. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J. Mol. Biol.* **333**, 721–745 (2003).
41. Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
42. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nature Methods* **11**, 63–65 (2014).
43. Baker, M. L. *et al.* Modeling protein structure at near atomic resolutions with Gorgon. *J. Struct. Biol.* **174**, 360–373 (2011).
44. Jiang, W., Baker, M. L., Ludtke, S. J. & Chiu, W. Bridging the information gap: computational tools for intermediate resolution structure interpretation. *J. Mol. Biol.* **308**, 1033–1044 (2001).
45. Cole, C., Barber, J. D. & Barton, G. J. The Jpred 3 secondary structure prediction server. *Nucleic Acids Res.* **36**, W197–W201 (2008).
46. Kelley, L. A. & Sternberg, M. J. Protein structure prediction on the Web: a case study using the Phyre server. *Nature Protocols* **4**, 363–371 (2009).
47. Abeyasinghe, S., Baker, M. L., Chiu, W. & Ju, T. Semi-isometric registration of line features for flexible fitting of protein structures. *Comput. Graph. Forum* **29**, 2243–2252 (2010).
48. Baker, M. L., Baker, M. R., Hryc, C. F. & Dimaio, F. Analyses of subnanometer resolution cryo-EM density maps. *Methods Enzymol.* **483**, 1–29 (2010).
49. Baker, M. L., Zhang, J., Ludtke, S. J. & Chiu, W. Cryo-EM of macromolecular assemblies at near-atomic resolution. *Nature Protocols* **5**, 1697–1708 (2010).
50. Baker, M. L., Baker, M. R., Hryc, C. F., Ju, T. & Chiu, W. Gorgon and pathwalking: macromolecular modeling tools for subnanometer resolution density maps. *Biopolymers* **97**, 655–668 (2012).
51. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
52. Baker, M. R., Rees, I., Ludtke, S. J., Chiu, W. & Baker, M. L. Constructing and validating initial C $\alpha$  models from subnanometer resolution density maps with pathwalking. *Structure* **20**, 450–463 (2012).
53. Li, Y. & Zhang, Y. REMO: A new protocol to refine full atomic protein models from C- $\alpha$  traces by optimizing hydrogen-bonding networks. *Proteins* **76**, 665–676 (2009).
54. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
55. de Beer, T. A., Berka, K., Thornton, J. M. & Laskowski, R. A. PDBSum additions. *Nucleic Acids Res.* **42**, D292–D296 (2014).
56. Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372**, 774–797 (2007).



**Extended Data Figure 1 | Cryo-EM of the purified InsP<sub>3</sub>R1.** **a**, Representative electron image of ice-embedded InsP<sub>3</sub>R1 taken at a defocus of 2.1  $\mu$ m and recorded using the K2 Summit camera. Shown image is motion-corrected. **b**, Fourier power spectrum of the micrograph shown in **a** with Thon rings

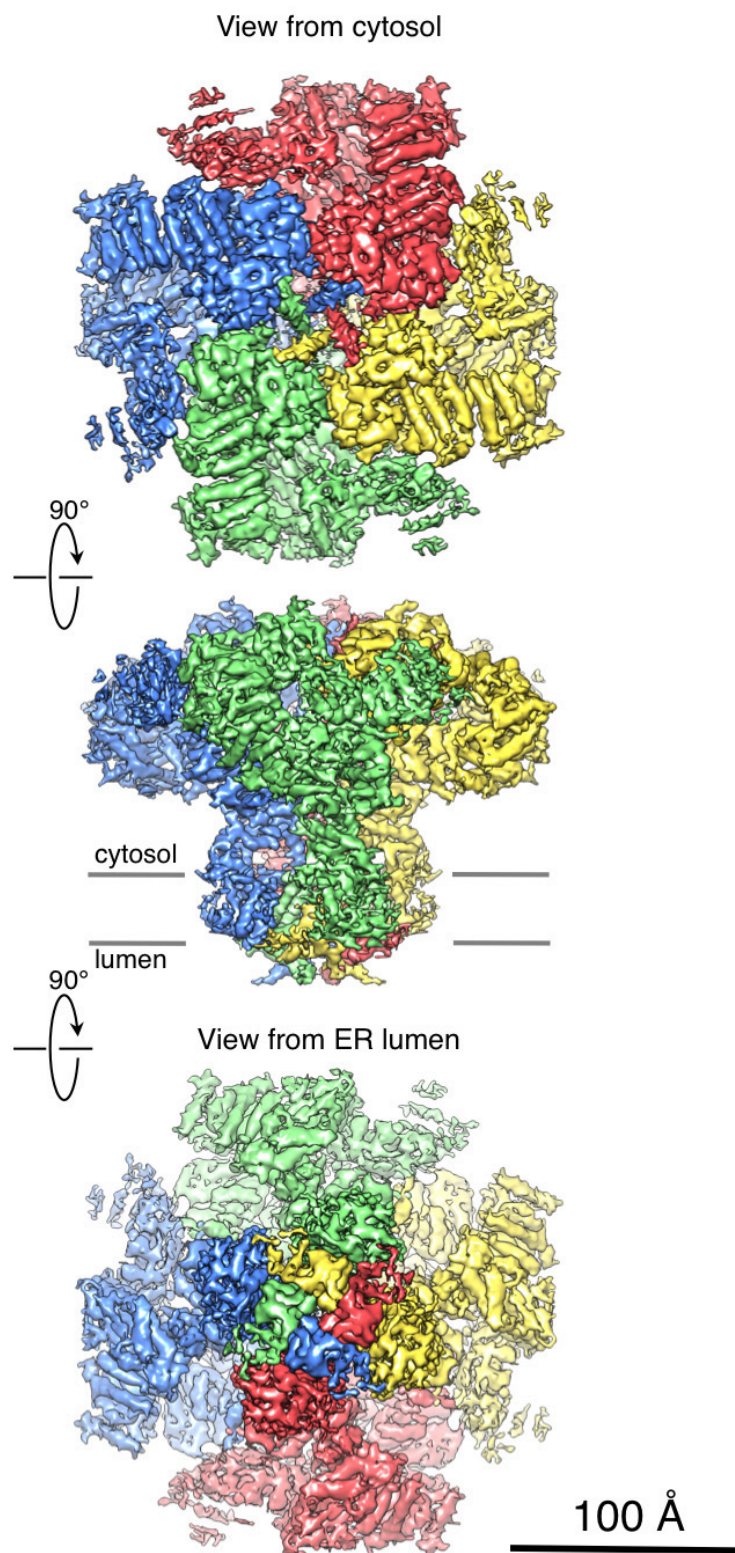
extending to 3.8 Å. **c**, 2D projections of the InsP<sub>3</sub>R1 map from the last iteration of refinement by RELION1.3 (rows 1 and 4) and EMAN2.1 (rows 2 and 5) are shown with corresponding 2D class averages (rows 3 and 6).



**Extended Data Figure 2 | Resolution estimation of cryo-EM 3D reconstruction.** **a**, The gold-standard FSC curve for the final cryo-EM 3D reconstruction generated with RELION1.3 (black) and EMAN2.1 (green). The overall resolution is 4.7 Å using the FSC cut-off = 0.143. **b**, Shown is the FSC curve between the cryo-EM density map generated with RELION1.3 and the molecular model (blue, FSC cut-off is 0.5) and the FSC curve between the 3D reconstructions generated with EMAN2.1 and RELION1.3 (red, FSC cut-off is 0.25). **c**, The cryo-EM density map of InsP<sub>3</sub>R1 generated with RELION1.3 is colour-coded based on ResMap (Methods). The map is depicted

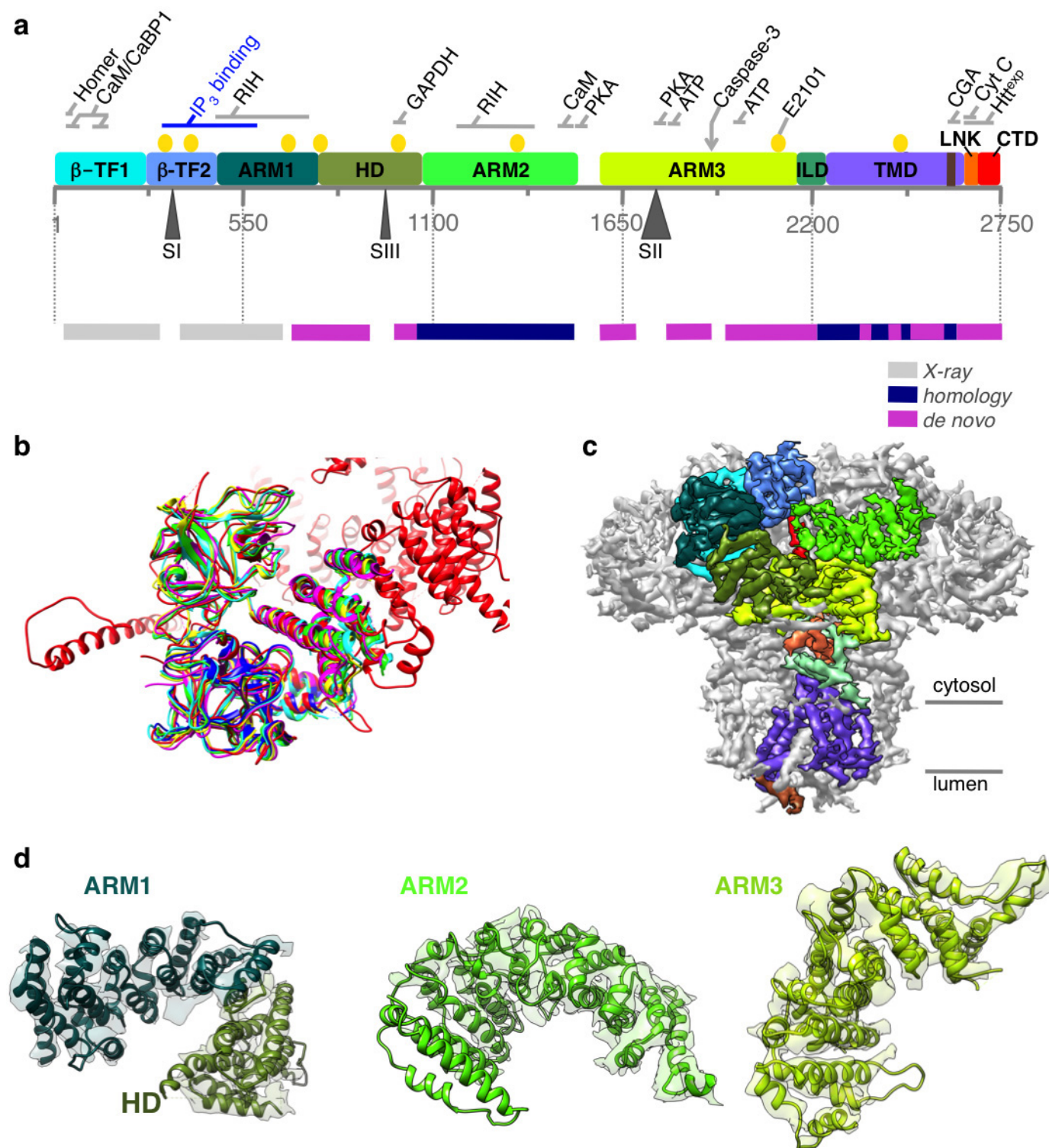
as viewed along the membrane plane (left) or as a slab of the density (dashed box in left panel) coincident with the four-fold channel axis (right). The cryo-EM density map exhibits local resolution variation ranging from 3.6 Å to 6.5 Å with the most highly resolved densities in the TMD, while parts of the peripheral densities in the cytoplasmic region are of lower resolution. The low resolution peripheral density in the transmembrane region is attributed to detergent bound to the protein. **d**, Euler angle distribution of all particles used for the final 3D reconstruction. Each view is represented by a sphere, for which the size is proportional to the number of particles for this specific view.





**Extended Data Figure 3 | 3D cryo-EM density map of the tetrameric InsP<sub>3</sub>R1 visualized in three orthogonal orientations.** InsP<sub>3</sub>R1 viewed from the cytosol (top), along the membrane plane (middle) and from the lumen

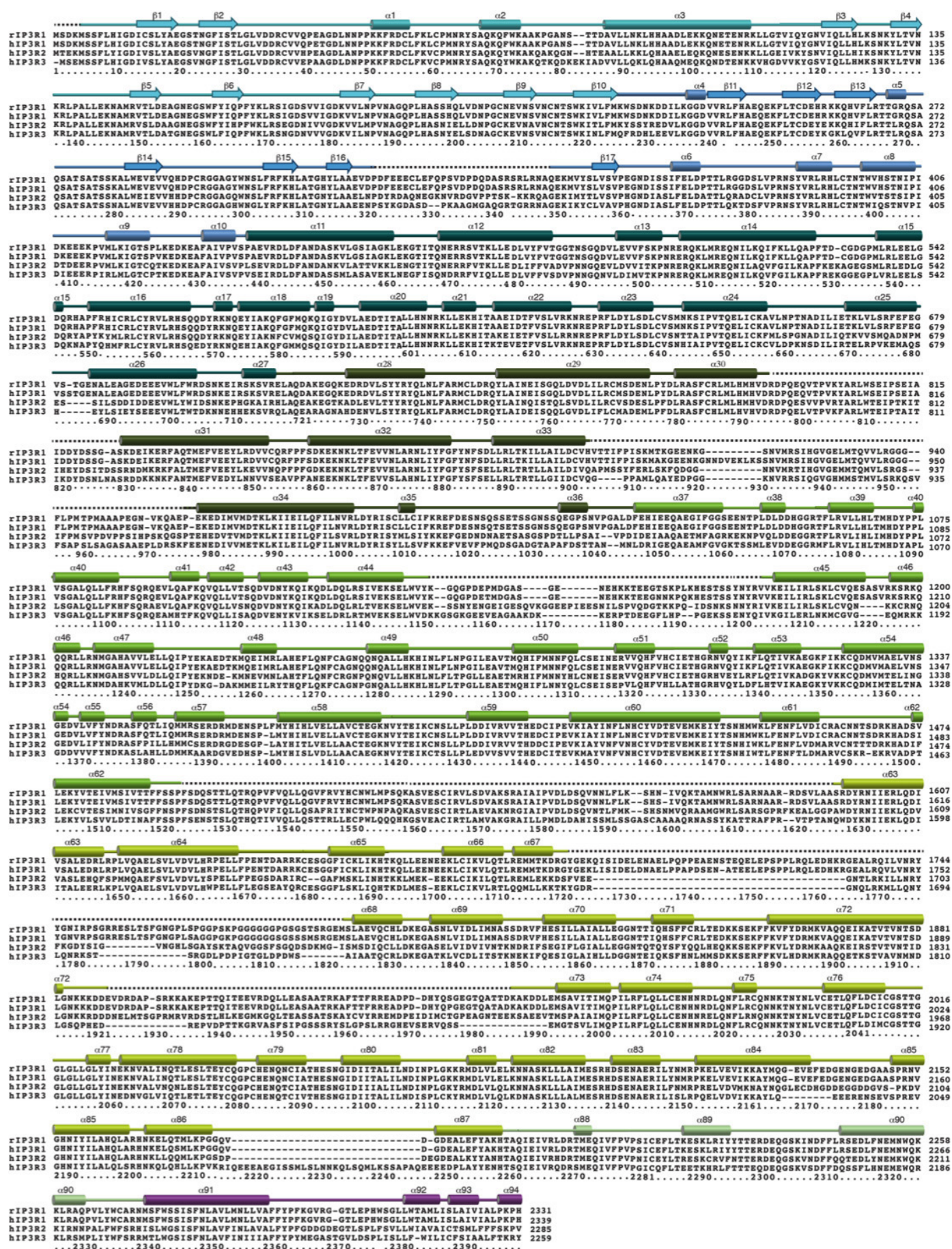
(bottom). Four individual subunits are colour-coded. The map is rendered at a threshold level corresponding to a molecular mass of ~1.3 MDa.



**Extended Data Figure 4 | Building an atomic model of InsP<sub>3</sub>R1.** **a**, Linear representation of a primary structure of InsP<sub>3</sub>R1 protein (GI accession 17380349). Ten domains identified in the cryo-EM density map are colour-coded. Three sites of alternative splicing (residues 318–332/SI, 918–926/SIII and 1692–1731/SII) are indicated below the sequence bar. Putative binding sites for several channel-specific ligands are indicated above the domains (ATP, ATP-binding CaM/CaBP, calmodulin/Ca<sup>2+</sup> binding protein; CGA, chromogranin A; cyt c, cytochrome c; Htt<sup>exp</sup>, huntingtin; PKA, protein kinase A; RIH, RyR/InsP<sub>3</sub>R homology; yellow circles denote Ca<sup>2+</sup>-binding<sup>20</sup>). The panel below shows a linear diagram of the protein sequence colour-coded based on the approach used for modelling different regions in the primary structure. The spaces between the bars correspond to unmodelled sequence (see also

Extended Data Fig. 5). **b**, The structure of the N-terminal domain (NTD) based on cryo-EM density map (red) is shown overlapped with the X-ray crystal structures of NTD (r.m.s.d. values = 1.3–1.4 Å): 1XZZ (blue); 3T8Sa (cyan); 3T8Sb (green); 3UJ4 (yellow); and 3UJ0 (magenta). **c**, Cryo-EM density map is viewed along the membrane plane. Densities corresponding to the individual domains of one subunit of InsP<sub>3</sub>R1 are colour-coded as in **a**. **d**, Models for solenoid-like  $\alpha$ -helical domains are shown superimposed on their corresponding cryo-EM densities. ARM1–ARM3, armadillo repeat domains 1–3; HD,  $\alpha$ -helical domain. The densities of ARM2 domain are less resolved than those of ARM1 and ARM3 but are sufficient to trace the backbone.



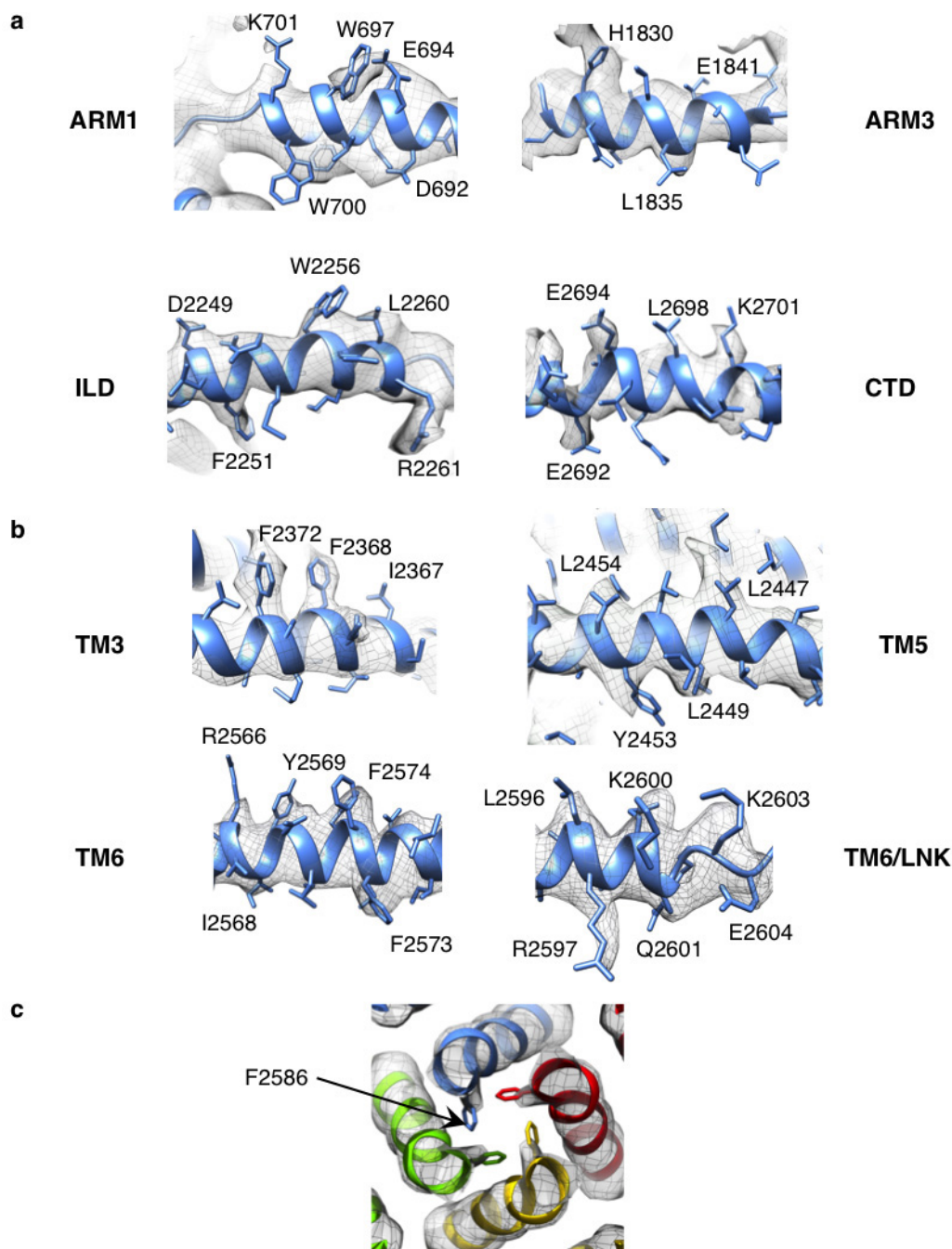




**Extended Data Figure 5 | Sequence alignment of selected InsP<sub>3</sub>R channels.**

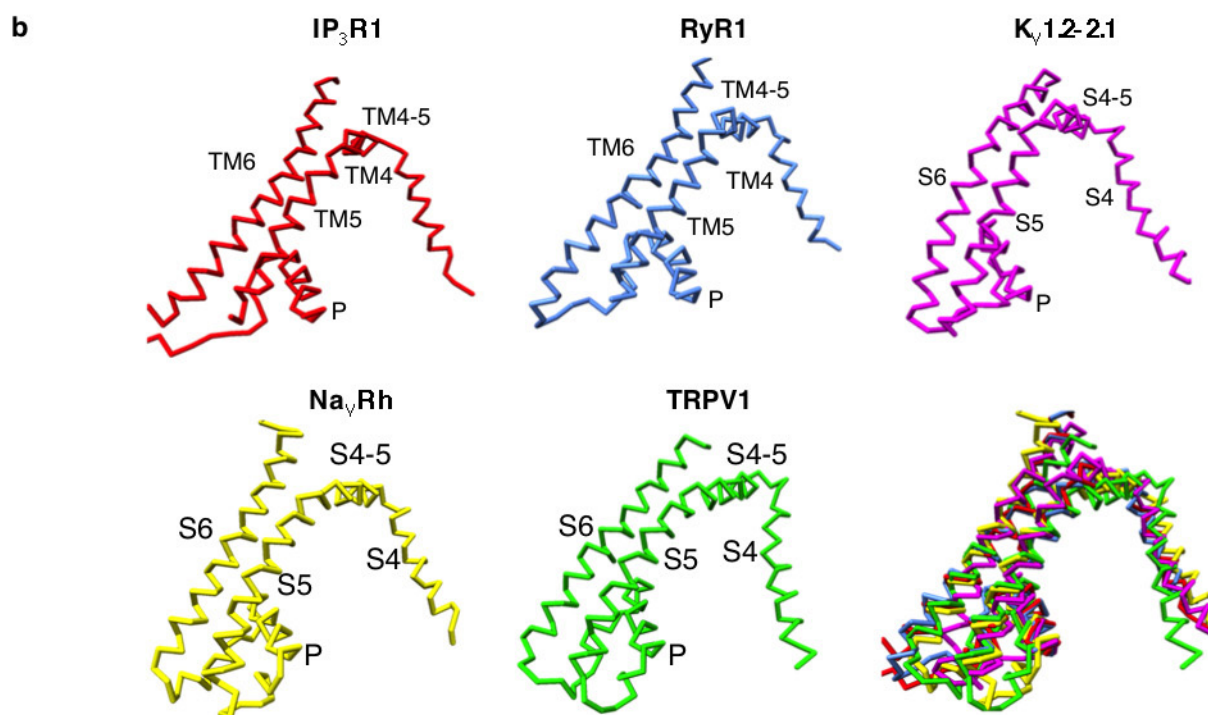
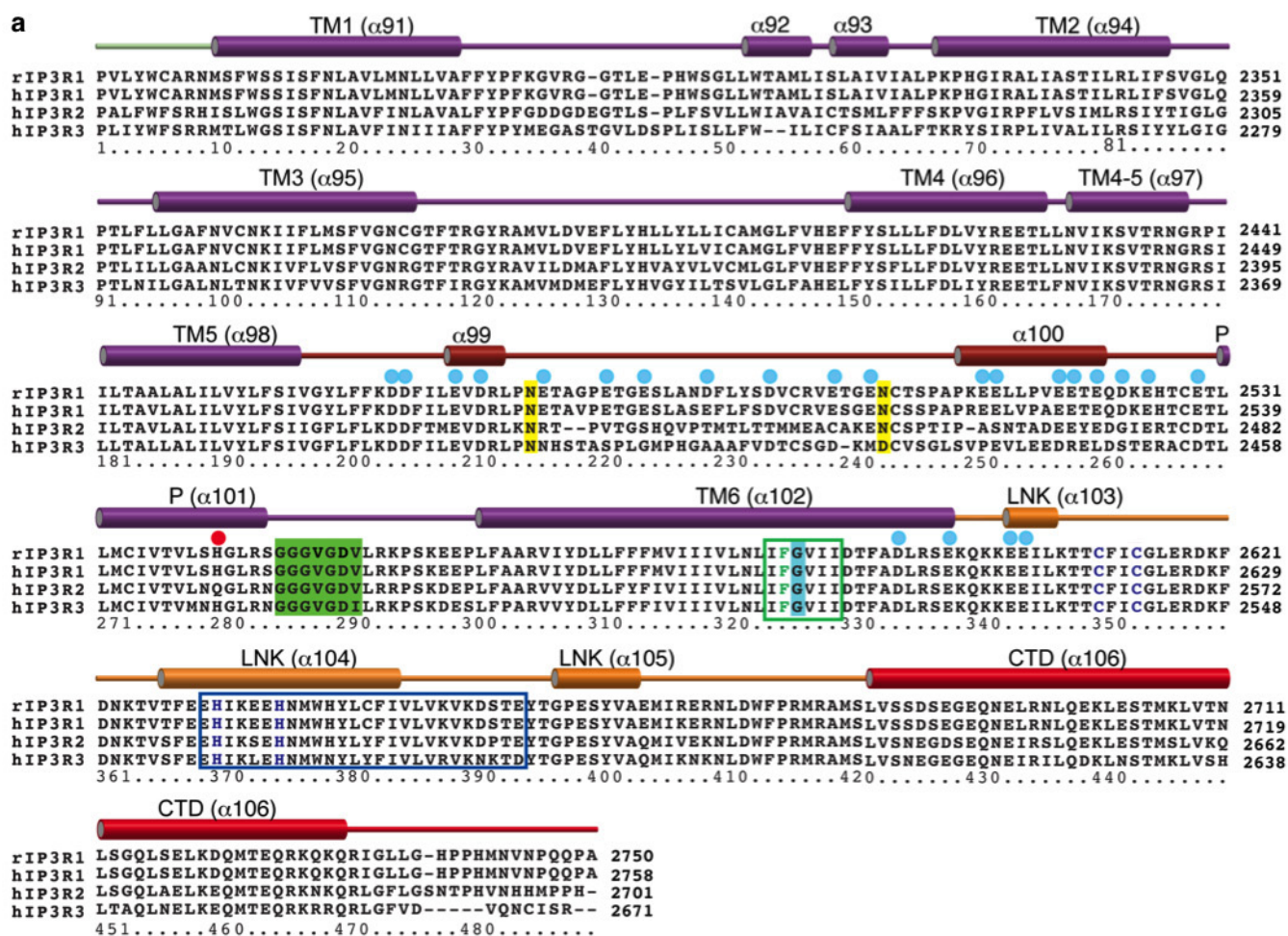
rInsP<sub>3</sub>R1, *Rattus norvegicus* (GI accession 17380349); hInsP<sub>3</sub>R1, *Homo sapiens* (GI accession 519668682); hInsP<sub>3</sub>R2, *Homo sapiens* (GI accession 259016258); hInsP<sub>3</sub>R3, *Homo sapiens* (GI accession 209572633); the primary sequence numbering includes the first methionine. The numbering of residues is given below the sequences, secondary structure elements are

indicated above the sequences and colour-coded in correspondence to the domains shown in Fig. 1b; dashed lines indicate regions that were not sufficiently resolved to be modelled. Given the enormous size of InsP<sub>3</sub>R proteins, the full-length sequence alignment was divided into two panels: sequence alignment for the transmembrane domains is shown in Extended Data Fig. 7a (note, overlap at the loop between the helices  $\alpha$ 90 and  $\alpha$ 91).



**Extended Data Figure 6 | Representative cryo-EM densities.** a–c, The cryo-EM densities for some regions are shown overlaid with the corresponding model: cytosolic helices (a); transmembrane helices (b); the constriction point

at Phe2586 within the ion conduction pathway (c), view from the cytosol along the four-fold axis, colour-coded by subunit.

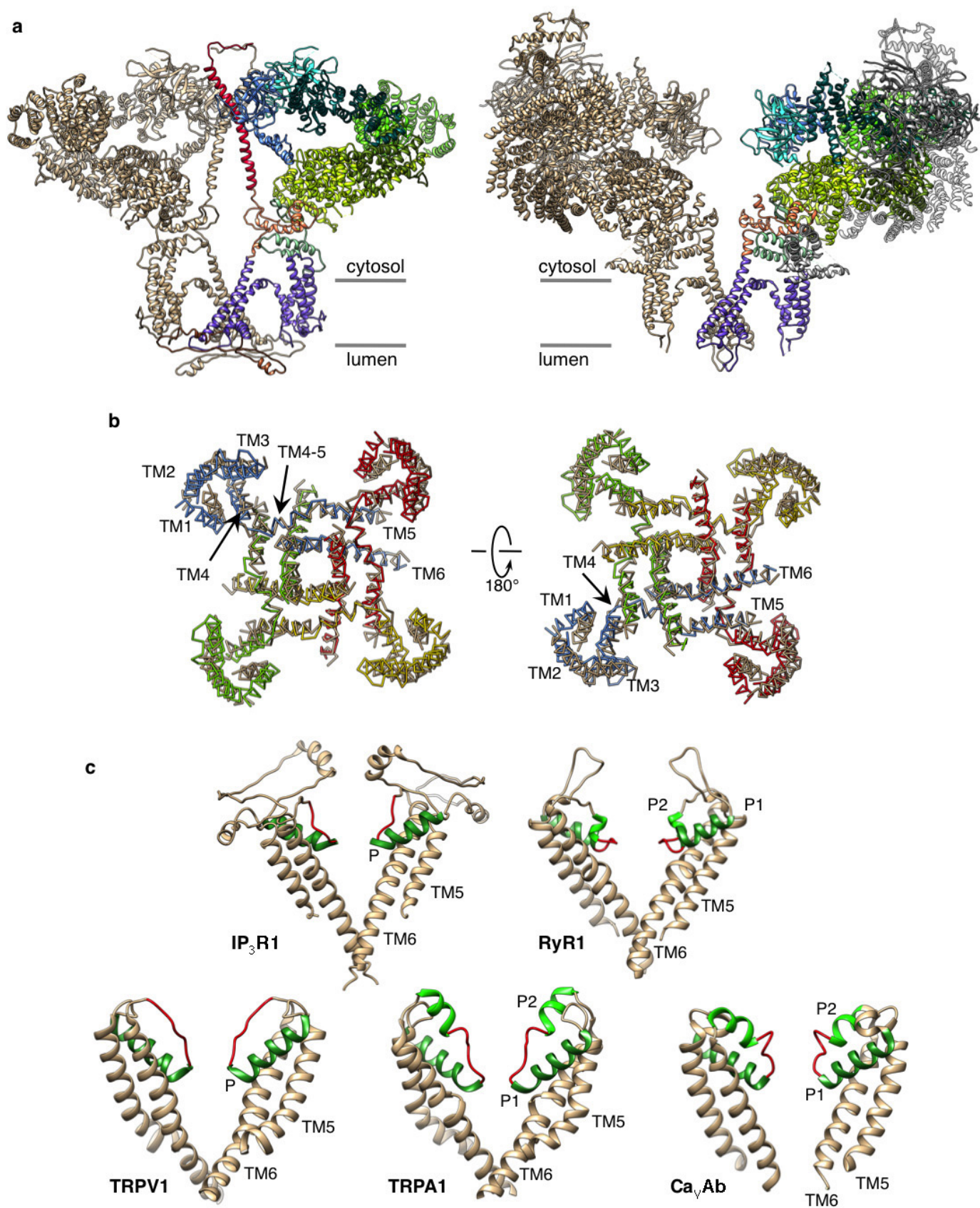




**Extended Data Figure 7 | Structural conservation of the pore among**

**tetrameric cation channels.** **a**, Alignment of the channel-forming domains; residues discussed in the text are labelled as following: blue circles, negatively charged residues; red circles, positively charged His2541; yellow highlight, *N*-glycosylation sites (Asn2476 and Asn2504); green highlight, selectivity filter; blue highlight, conserved Gly2587; green box, hydrophobic constriction region

and Phe2586 shown in green; dark blue, Zn<sup>2+</sup>-finger-like residues Cys2611/Cys2614 and His2631/His2636; blue box, tetramerization region. **b**, Structural comparison of pore-forming elements of InsP<sub>3</sub>R1, RyR1, K<sub>v</sub>1.2–2.1, Na<sub>v</sub>Rh, TRPV1 (PDB accessions: 3J8H, 2R9R, 4DXW and 3J5P, respectively). Note substantial overlap between structures.

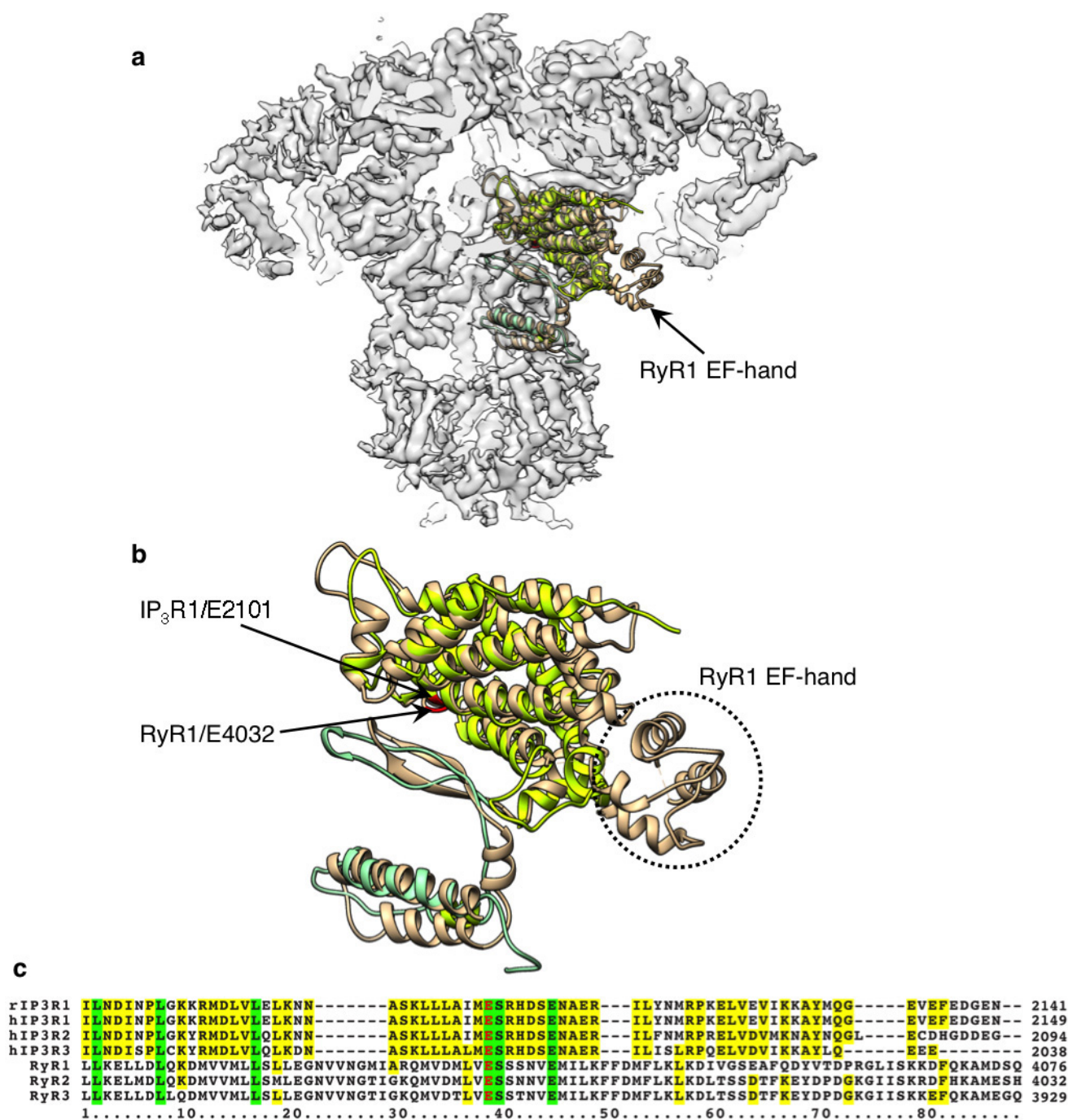


**Extended Data Figure 8 | Comparison of InsP<sub>3</sub>R1 and RyR1 structures.**

**a**, Two opposing subunits of InsP<sub>3</sub>R1 and RyR1 (PDB accession 3J8H) are viewed along the membrane plane. One InsP<sub>3</sub>R1 subunit is colour-coded by domain (left). Structurally consistent domains in one RyR1 subunit are colour-coded using InsP<sub>3</sub>R1 domain architecture. Domains of RyR1 not in common are shown in grey. **b**, TMDs of RyR1 (tan) and InsP<sub>3</sub>R1 (coloured by subunit) are superimposed using Chimera's Matchmaker and viewed from the

cytosol (left) and lumen (right). The r.m.s.d. between 80 atom pairs is 2.0 Å. For clarity, P-helices are not shown. **c**, Structural comparison of the selectivity filter (red) in InsP<sub>3</sub>R1 with that in some tetrameric cation channels: RyR1, TRPV1, TRPA1, Ca<sub>v</sub>Ab (PDB accessions: 3J8H, 3J5Q, 3J9P and 4MS2, respectively). Two opposing subunits are shown; TM5 and TM6 helices are coloured tan, P-helices are in green.

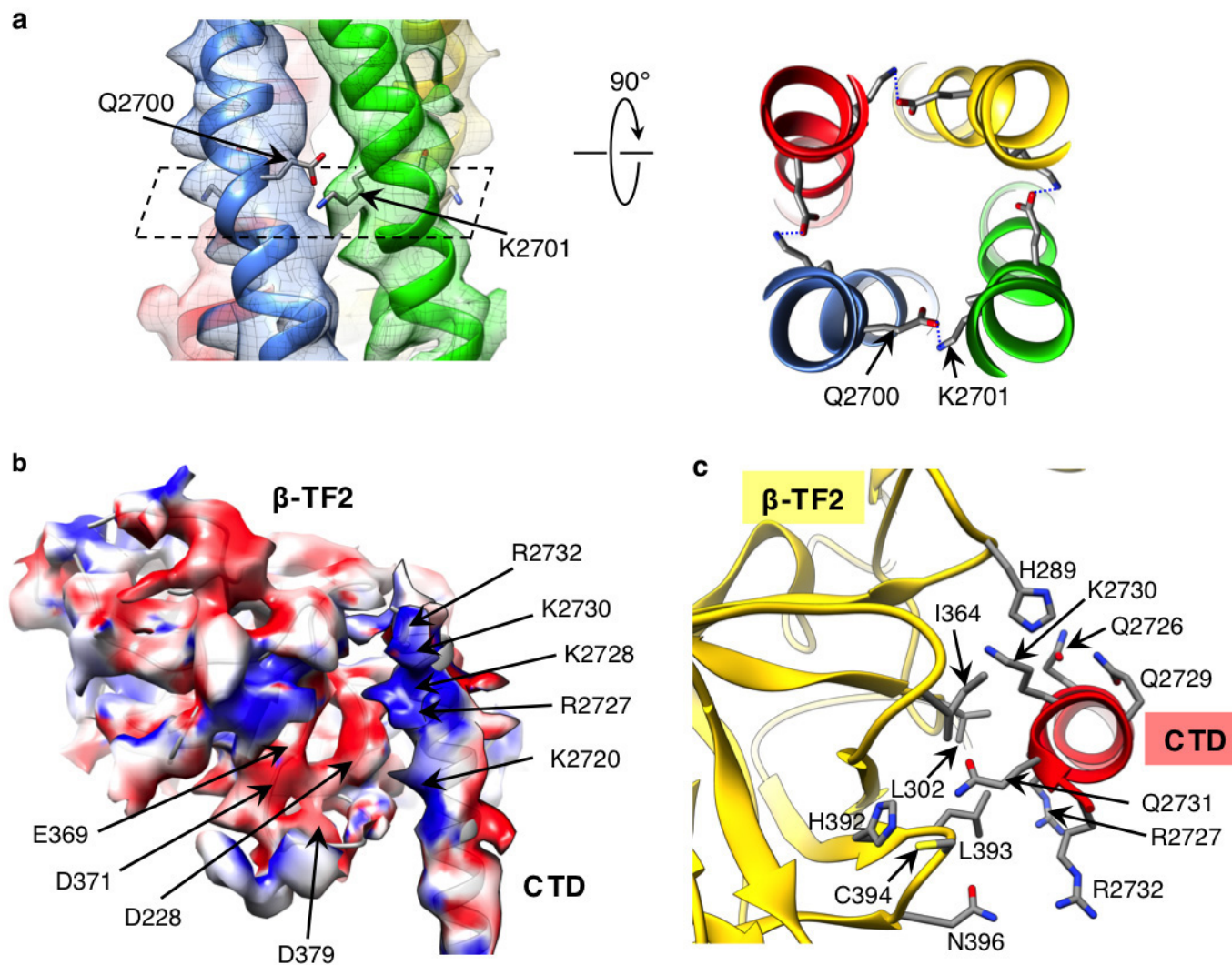




### Extended Data Figure 9 | The putative cytosolic Ca<sup>2+</sup> sensor in InsP<sub>3</sub>R1.

**a**, Cut-open side view of the cryo-EM density map of InsP<sub>3</sub>R1 is shown with the structures of the Ca<sup>2+</sup> sensor region for InsP<sub>3</sub>R1 (residues 1952–2270) and RyR1 (residues 3877–4251; PDB accession 3J8H); the EF-hand of RyR1 includes residues 4071–4130. **b**, Close-up view of the overlapped Ca<sup>2+</sup> sensor structures for InsP<sub>3</sub>R1 (colour-coded by domain as in Fig. 1) and RyR1 (tan). The conserved Glu2101/InsP<sub>3</sub>R1 and Glu4032/RyR1 are shown in red.

**c**, Sequence alignment of the predicted Ca<sup>2+</sup> sensor regions comprising the conserved Glu2101/InsP<sub>3</sub>R1 and Glu4032/RyR1 (red) (rInsP<sub>3</sub>R1/GI code:17380349; hInsP<sub>3</sub>R1/GI code: 519668682; hInsP<sub>3</sub>R2/GI code: 259016258; hInsP<sub>3</sub>R3/GI code: 209572633; RyR1/GI code: 134134; RyR2/GI code: 308153559; RyR3/GI code: 75074791); green highlight denotes completely conserved residues; yellow highlight denotes identical residues.



**Extended Data Figure 10 | Structural coupling in the CTD.** **a**, A bundle of the CTD helices is viewed perpendicular to the channel axis at the position of a predicted inter-subunit saltbridge between residues Gln2700 and Lys2701. The right panel shows the predicted salt bridge normal to the channel axis. Helices are colour-coded per subunit. **b**, Structures of the CTD and  $\beta$ -TF2 domains are superimposed on their corresponding cryo-EM densities and

the surfaces are colour-coded according to electrostatic charges calculated for the model: blue denotes positive charges; red denotes negative charges. Shown is a side view with the cytoplasmic side facing up. **c**, Close-up view of the interface between the CTD and  $\beta$ -TF2 domains of two neighbouring subunits; viewed from cytosol perpendicular to the membrane plane.

# Accreting protoplanets in the LkCa 15 transition disk

S. Sallum<sup>1</sup>, K. B. Follette<sup>1,2</sup>, J. A. Eisner<sup>1</sup>, L. M. Close<sup>1</sup>, P. Hinz<sup>1</sup>, K. Kratter<sup>1</sup>, J. Males<sup>1</sup>, A. Skemer<sup>1</sup>, B. Macintosh<sup>2</sup>, P. Tuthill<sup>3</sup>, V. Bailey<sup>1</sup>, D. Defrère<sup>1</sup>, K. Morzinski<sup>1</sup>, T. Rodigas<sup>4</sup>, E. Spalding<sup>1</sup>, A. Vaz<sup>1</sup> & A. J. Weinberger<sup>4</sup>

Exoplanet detections have revolutionized astronomy, offering new insights into solar system architecture and planet demographics. While nearly 1,900 exoplanets have now been discovered and confirmed<sup>1</sup>, none are still in the process of formation. Transition disks, protoplanetary disks with inner clearings<sup>2–4</sup> best explained by the influence of accreting planets<sup>5</sup>, are natural laboratories for the study of planet formation. Some transition disks show evidence for the presence of young planets in the form of disk asymmetries<sup>6,7</sup> or infrared sources detected within their clearings, as in the case of LkCa 15 (refs 8, 9). Attempts to observe directly signatures of accretion onto protoplanets have hitherto proven unsuccessful<sup>10</sup>. Here we report adaptive optics observations of LkCa 15 that probe within the disk clearing. With accurate source positions over multiple epochs spanning 2009–2015, we infer the presence of multiple companions on Keplerian orbits. We directly detect H $\alpha$  emission from the innermost companion, LkCa 15 b, evincing hot (about 10,000 kelvin) gas falling deep into the potential well of an accreting protoplanet.

We observed LkCa 15 using the high-contrast imaging technique of non-redundant masking (NRM)<sup>11</sup>, at the Large Binocular Telescope (LBT) in Ks ( $\lambda_c = 2.16 \mu\text{m}$ ) and L' ( $\lambda_c = 3.7 \mu\text{m}$ ; see Extended Data Table 1). We detect two components, LkCa 15 b and c, in both bands, with consistent positions across wavelength given the uncertainties (see Table 1, Extended Data Fig. 1). We detect a faint, third component, LkCa 15 d, at L' only. Since d is significantly fainter than b and c, and not detected at Ks, we focus on the other two sources in the following analysis, but include discussion of the putative third companion where relevant.

We also observed LkCa 15 in H $\alpha$  ( $\lambda_c = 655.8 \text{ nm}$ ) using the Magellan Adaptive Optics System (MagAO) in Simultaneous Differential Imaging (SDI)<sup>12,13</sup> mode (see Methods). We detect LkCa 15 b in these data, at a signal-to-noise of 6.4 and a position that agrees with the LBT observations (see Table 1, Extended Data Figs 2–5, Extended Data Table 2). LkCa 15 c was not detected in H $\alpha$ , perhaps owing to higher extinction along the line of sight or lower accretion rates at the time of the observations. Both b and c lie well within the disk clearing (Fig. 1), which extends to a stellocentric radius of  $56 \text{ AU}^{14}$ .

We compare the positions of LkCa 15 b and c to the infrared signal seen in 2009–2010 NRM observations<sup>8</sup>. As shown in Fig. 2, orbital fits (fixed to the outer disk plane: inclination  $i = 50^\circ$ , position angle  $\theta = 150^\circ$ )<sup>14</sup> suggest distinct orbits, with b moving faster (semimajor axis,  $a = 14.7 \pm 2.1 \text{ AU}$ ) than c ( $a = 18.6 \pm 2.5 \text{ AU}$ ). Taking the semimajor axis uncertainties into account and requiring that these orbits be stable, b and c must have masses lower than 5–10 times that of Jupiter ( $M_J$ )<sup>15,16</sup>, with masses over  $5 M_J$  allowed only in the case of a 2:1 resonance. For completeness, we performed a series of four-body simulations to show that stable orbital solutions exist including LkCa 15 d, with three planet masses  $\leq 0.5 M_J$  (see Methods, Extended Data Figs 6, 7) and higher masses for b and c allowed with a less massive d.

We calculate infrared fluxes for LkCa 15 b and c, and compare them to circumplanetary accretion disk models<sup>17,18</sup> and hot-start<sup>19</sup> models of sub-stellar mass companions shortly after accretion has ceased (see Fig. 3). From the LkCa 15 A magnitudes (2MASS Ks =  $8.16^{20}$  and IRAC  $m_{3.6} = 7.61^{21}$ ), we derive fluxes of  $1.4 \pm 0.7 \text{ mJy}$  at Ks and  $2.5 \pm 1.2 \text{ mJy}$

**Table 1 | Model and experimental results**

Component	Date	Instrument	$\lambda$	PA* (°)	s† (mas)	$\Delta\ddagger$ (mag)	M§ (mag)	$M_p\dot{M}  $ ( $M_J^2 \text{ yr}^{-1}$ )	a¶ (AU)
Model fit results									
LkCa 15 b	15 Nov 2014	MagAO	H $\alpha$ = 656.3 nm	$-104 \pm 3$	$93 \pm 8$	$5.2 \pm 0.3$	$15.8 \pm 0.3$	$3 \times 10^{-6}$	$14.7 \pm 2.1$
LkCa 15 b	5–7 Feb 2015	LBT	Ks = $2.18 \mu\text{m}$	$-86 \pm 26_{16}^{26}$	$125 \pm 25_{40}^{25}$	$6.0 \pm 2.0_{0.5}^{2.0}$	$14.2 \pm 2.0_{0.5}^{2.0}$	$10^{-5}$	$14.7 \pm 2.1$
LkCa 15 b	15 Dec 2014	LBT	L' = $3.8 \mu\text{m}$	$-100 \pm 21_{16}^{21}$	$106 \pm 81_{19}^{81}$	$5.4 \pm 0.1_{4.9}^{0.1}$	$13.6 \pm 0.1_{4.9}^{0.1}$	$10^{-5}$	$14.7 \pm 2.1$
LkCa 15 c	5–7 Feb 2015	LBT	Ks = $2.18 \mu\text{m}$	$-48 \pm 22_{10}^{22}$	$85 \pm 15_{15}^{15}$	$5.5 \pm 0.5_{0.5}^{0.5}$	$13.7 \pm 0.5_{0.5}^{0.5}$	$10^{-5}$	$18.6 \pm 2.5_{2.7}^{2.5}$
LkCa 15 c	15 Dec 2014	LBT	L' = $3.8 \mu\text{m}$	$-44 \pm 16_{21}^{16}$	$68 \pm 37_{43}^{37}$	$4.8 \pm 0.7_{4.3}^{0.7}$	$12.9 \pm 0.7_{4.3}^{0.7}$	$10^{-5}$	$18.6 \pm 2.5_{2.7}^{2.5}$
LkCa 15 d	15 Dec 2014	LBT	L' = $3.8 \mu\text{m}$	$14 \pm 32_{24}^{32}$	$87 \pm 72_{70}^{72}$	$5.9 \pm 2.1_{5.4}^{2.1}$	$14.1 \pm 2.1_{5.4}^{2.1}$	$5 \times 10^{-6}$	$18.0 \pm 6.7_{5.4}^{6.7}$
LBT joint fit results									
LkCa 15 b	Dec 2014–Feb 2015	LBT	Ks + L'	$-98 \pm 12_{10}^{12}$	$95 \pm 50_{15}^{50}$	$\Delta Ks = 6.0 \pm 0.5$ $\Delta L' = 5.0 \pm 0.5$	$14.2 \pm 0.5$ $13.2 \pm 0.5$	$10^{-5}$	$14.7 \pm 2.1$
LkCa 15 c	Dec 2014–Feb 2015	LBT	Ks + L'	$-42 \pm 12_{10}^{12}$	$80 \pm 15_{10}^{15}$	$\Delta Ks = 5.5 \pm 0.5$ $\Delta L' = 5.0 \pm 0.5$	$13.7 \pm 0.5$ $13.2 \pm 0.5$	$10^{-5}$	$18.6 \pm 2.5_{2.7}^{2.5}$

\*Position angle measured east of north.

†Stellocentric separation.

‡Contrast.

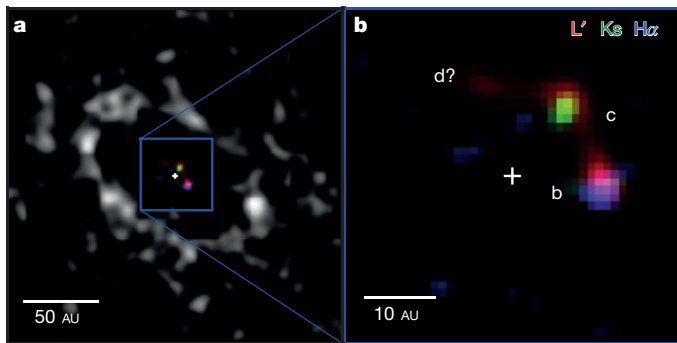
§Absolute magnitude.

|| Planet mass times accretion rate.

¶Best fit orbital semi-major axis.

<sup>1</sup>Astronomy Department, University of Arizona, 933 North Cherry Avenue, Tucson, Arizona 85721, USA. <sup>2</sup>Kavli Institute for Particle Astrophysics and Cosmology, Stanford University, Stanford, California 94305, USA. <sup>3</sup>School of Physics, University of Sydney, Sydney, New South Wales 2006, Australia. <sup>4</sup>Department of Terrestrial Magnetism, Carnegie Institution for Science, 5241 Broad Branch Rd NW, Washington, Washington DC 20015, USA.



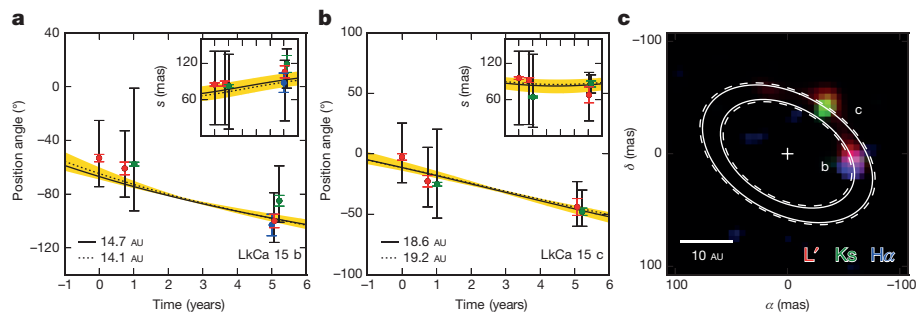


**Figure 1 | Composite H $\alpha$ , Ks, and L' image.** **a**, The coloured image shows H $\alpha$  (blue), Ks (green), and L' (red) detections at the same scale as VLA millimetre observations<sup>29</sup> (greyscale). **b**, Zoomed in composite image of LBT and Magellan observations, with b, c, and d marked.

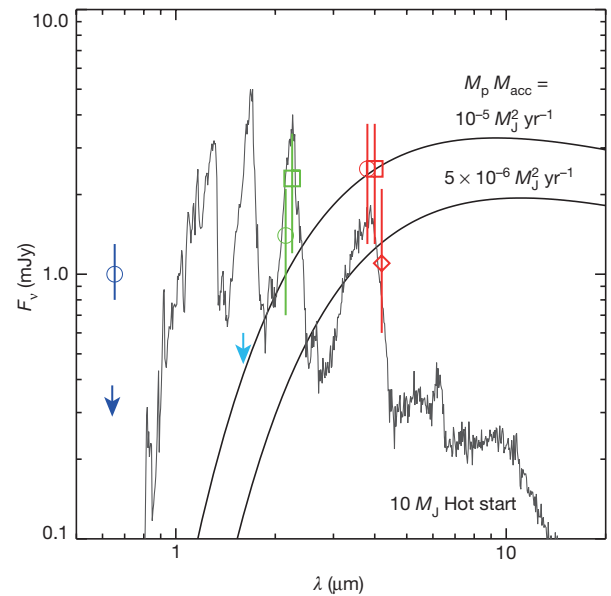
at L' for b, and  $2.3 \pm 1.1$  mJy at Ks and  $2.5 \pm 1.2$  mJy at L' for c. These are consistent with accretion disks having inner radii  $R_{\text{in}} = 2R_J$  and planet mass times accretion rate  $M_p \dot{M} \approx 10^{-5} M_J^2 \text{ yr}^{-1}$ . However, changing  $R_{\text{in}}$  affects both the total disk flux and its colour. The large uncertainties on fluxes and colours allow us to constrain  $R_{\text{in}}$  only to within a factor of  $\sim 2$ , translating to a factor of  $\sim 2$ – $3$  uncertainty in  $M_p \dot{M}$  (for example, a  $R_{\text{in}} = 1R_J$ ,  $M_p \dot{M} \approx 3 \times 10^{-6} M_J^2 \text{ yr}^{-1}$  disk can also reproduce the observations).

While the hot-start model shown in Fig. 3 can approximately produce the Ks and L' emission for b and c, the observations are best explained by an accretion disk model. The hot start model can only match a previously established  $1.55 \mu\text{m}$  upper limit on the contrast of the structure within the disk gap ( $\Delta H = 7.2 \text{ mag}$ )<sup>9</sup> if the extinction is significantly higher than inferred towards the star. Moreover, even a highly extinguished hot-start model cannot reproduce the strong emission at  $4.7 \mu\text{m}$  (contrast of  $\Delta M = 3.5 \text{ mag}$ )<sup>9</sup>. Emission from an accretion disk increases from L' to M band, while the hot-start model produces little M band emission. Finally, a cooling photosphere produces no H $\alpha$  emission, firmly ruling out the hot-start model as the source of LkCa 15 b.

Since LkCa 15 b is detected at H $\alpha$ , an accretion tracer<sup>22–24</sup>, its nature as an accreting protoplanet is clear. LkCa 15 b's H $\alpha$  contrast, corrected for A's H $\alpha$  excess and assuming equal extinction to A ( $A_R = 0.75 \text{ mag}$ )<sup>25</sup>, corresponds to a line flux of  $\sim 6 \times 10^{-5} L_{\odot}$ . Assuming similar accretion luminosity ( $L_{\text{acc}}$ ) scalings as low-mass T Tauri stars<sup>12,23</sup> gives  $L_{\text{acc}} \approx 4 \times 10^{-4} L_{\odot}$ , yielding  $M_p \dot{M} \approx 3 \times 10^{-6} M_J^2 \text{ yr}^{-1}$  for a  $1.6R_J$  planet<sup>26</sup> ( $R_J$ , Jupiter radius). Previous observations showed that low-mass, accreting objects may emit a higher fraction of accretion luminosity at H $\alpha$ <sup>22</sup>; assuming similar accretion scalings as T Tauri stars may overestimate  $L_{\text{acc}}$ . Extinction towards b is also uncertain; while we assume equal extinction to A and b, localized extinction can alter the numbers quoted above. While the uncertainties are large, this  $M_p \dot{M}$  is consistent with that estimated from the infrared fluxes.



**Figure 2 | Position evolution.** **a**, LkCa 15 b position angle and separation (inset) evolution, showing H $\alpha$  (blue), Ks (green), and L' (red). The earliest three points indicate previous observations<sup>8</sup>; others show fits to our data. Coloured and black  $1\sigma$  error bars are from a nonlinear algorithm and a grid, respectively (see Methods). The yellow shading spans the  $1\sigma$  allowed



**Figure 3 | Spectral energy distributions.** Symbols indicate fluxes for LkCa 15 b (circles), c (squares), and d (diamonds), showing H $\alpha$  (dark blue), Ks (green), and L' (red). The light and dark blue arrows mark previously-published H-band<sup>9</sup> and  $3\sigma$   $642 \text{ nm}$  upper limits for LkCa 15 b, respectively. The lines show accretion disk and hot-start models. The disk models are simple combinations of blackbody spectra<sup>17</sup>, a suitable approximation for the case of a cool ( $T < 1,500 \text{ K}$ ) stellar atmosphere where dust opacity dominates. The  $M_p \dot{M}$  calculated from the H $\alpha$  flux agrees with that inferred from the infrared measurements (see text).

Previous investigators posited a single protoplanet in LkCa 15, accreting material from its co-orbital surroundings<sup>8</sup>. While the semi-major axis uncertainties do not formally rule out b and c (and d, see Extended Data Fig. 6) being co-orbital, physical arguments show that they cannot be gravitationally bound. The size of the previously reported emission (several au) is larger than a Hill radius ( $\sim 1.8 \text{ au}$  for a  $10 M_J$  planet ( $M_J$ , Jupiter mass) orbiting a  $1 M_{\odot}$  star at  $10 \text{ au}$ ), and much larger than the maximum possible size of a circumplanetary disk ( $\sim 1/3$  the Hill radius<sup>27</sup>). Thus the sources cannot be part of a bound, accreting system, and an alternative scenario is required to explain the observations.

We argue further that it is difficult to explain LkCa 15 b and c (and d) with an orbiting clump of gravitationally unbound dust within the disk gap, emitting thermally or in scattered light. At a distance of  $\sim 10 \text{ au}$ , neither LkCa 15 A nor a companion with a contrast of  $\sim 5$  magnitudes can heat dust sufficiently to emit at  $2$ – $4 \mu\text{m}$ . Assuming isotropic single scattering, we calculate that an optically thin spherical clump of dust, perhaps resulting from a recent planetesimal collision, could produce the contrast observed at a single wavelength.

parameters from orbital fitting. Solid and dotted curves show stable orbits for  $0.5 M_J$  and  $1.0 M_J$  planets, respectively. **b**, Same as **a**, for LkCa 15 c. **c**, Stable orbits for  $0.5 M_J$  (solid) and  $1.0 M_J$  (dotted) planets. mas, milliarcsecond.

However, observing this clump before it sheared out would be a priori unlikely, since the viscous timescale at 10 AU is just  $\sim 3\%$  the age of the system.

Observations argue strongly against this explanation as well. Scattering cannot cause increasing emission from H to M band<sup>9</sup>, since dust opacity decreases with increasing wavelength. Furthermore, since dust opacity is equal between H $\alpha$  and the nearby continuum, scattering signals have equal contrast in both narrowband filters. Scaling the continuum image by the LkCa 15 A H $\alpha$ -to-continuum flux ratio and subtracting it from the H $\alpha$  image should only lead to an H $\alpha$  detection if scattering is not the emission mechanism. Indeed, this yields a LkCa 15 b detection with signal-to-noise of 4.8. While the Wollaston beam splitter in MagAO's SDI mode could lead to contamination by polarized light, the visible polarized scattered light intensity at b's position is less than  $\sim 7\%$  the H $\alpha$  source flux<sup>28</sup>. It could not cause the H $\alpha$  detection. This leaves the multiple-planet scenario as the most natural explanation for the data.

Both the infrared and H $\alpha$  observations show that we are unambiguously witnessing planet formation in LkCa 15. The data offer evidence that giant protoplanets undergo a period of high infrared and H $\alpha$  luminosity during their accretion phase. In the near future, ALMA's sensitivity and angular resolution should enable us to detect sub-millimetre emission from circumplanetary disks<sup>29</sup>. Additionally, while the LBT data published here were taken in single-aperture mode (baselines up to  $\sim 8$  m), non-redundant masking using the co-phased LBTI will provide 23-m baselines, allowing us to place tight constraints on the companion orbits and to resolve structure at smaller separations. Continued monitoring of accretion tracers from LkCa 15 b will probe whether the accretion is steady or stochastic. This young system provides the first opportunity to study planet formation and disk–planet interactions directly.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 16 May; accepted 10 September 2015.**

- Akeson, R. L. *et al.* The NASA Exoplanet Archive: data and tools for exoplanet research. *Publ. Astron. Soc. Pacif.* **125**, 989–999 (2013).
- Andrews, S. M. *et al.* Resolved images of large cavities in protoplanetary transition disks. *Astrophys. J.* **732**, 42–66 (2011).
- Strom, K. M., Strom, S. E., Edwards, S., Cabrit, S. & Skrutskie, M. F. Circumstellar material associated with solar-type pre-main-sequence stars - a possible constraint on the timescale for planet building. *Astron. J.* **97**, 1451–1470 (1989).
- Calvet, N. *et al.* Disks in transition in the Taurus population: Spitzer IRS spectra of GM Aurigae and DM Tauri. *Astrophys. J.* **630**, L185–L188 (2005).
- Bryden, G., Chen, X., Lin, D. N. C., Nelson, R. P. & Papaloizou, J. C. B. Tidally induced gap formation in protostellar disks: gap clearing and suppression of protoplanetary growth. *Astrophys. J.* **514**, 344–367 (1999).
- Isella, A. *et al.* An azimuthal asymmetry in the LkHa 330 disk. *Astrophys. J.* **775**, 30–40 (2013).
- Pérez, L. M., Isella, A., Carpenter, J. M. & Chandler, C. J. Large-scale asymmetries in the transitional disks of SAO 206462 and SR 21. *Astrophys. J.* **783**, L13–L18 (2014).
- Kraus, A. L. & Ireland, M. J. LkCa 15: a young exoplanet caught at formation? *Astrophys. J.* **745**, 5–16 (2012).
- Ireland, M. J. & Kraus, A. L. Orbital Motion and Multi-Wavelength Monitoring of LkCa15 b. In Booth, M., Matthews, B. C. & Graham, J. R. (eds) *Exploring the Formation and Evolution of Planetary Systems*, Vol. 299 of *IAU Symposium*, 199–203 (2014).
- Whelan, E. T. *et al.* Spectro-astrometry of LkCa 15 with X-Shooter: searching for emission from LkCa 15b. *J. Astron. Astrophys.* **579**, A48 (2015).
- Tuthill, P. G., Monnier, J. D. & Danchi, W. C. Aperture masking interferometry on the Keck I Telescope: new results from the diffraction limit. In Léna, P. & Quirrenbach, A. (eds) *Interferometry in Optical Astronomy*, Vol. 4006 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, 491–498 (2000).
- Close, L. M. *et al.* Discovery of H $\alpha$  emission from the close companion inside the gap of transitional disk HD 142527. *Astrophys. J.* **781**, L30–L34 (2014).
- Marois, C., Nadeau, D., Doyon, R., Racine, R. & Walker, G. A. H. Differential simultaneous imaging and faint companions: TRIDENT first results from CFHT. In Martin, E. (ed.) *Brown Dwarfs*, Vol. 211 of *IAU Symposium*, 275–278 (2003).
- Thalmann, C. *et al.* The architecture of the LkCa 15 transitional disk revealed by high-contrast imaging. *J. Astron. Astrophys.* **566**, A51 (2014).
- Gladman, B. Dynamics of systems of two close planets. *Icarus* **106**, 247–263 (1993).
- Beaugé, C., Ferraz-Mello, S. & Michtchenko, T. A. Extrasolar planets in mean-motion resonance: apses alignment and asymmetric stationary solutions. *Astrophys. J.* **593**, 1124–1133 (2003).
- Eisner, J. A. Spectral energy distributions of accreting protoplanets. *Astrophys. J.* **803**, L4–L8 (2015).
- Zhu, Z. Accreting circumplanetary disks: observational signatures. *Astrophys. J.* **799**, 16–24 (2015).
- Spiegel, D. S. & Burrows, A. Spectral and photometric diagnostics of giant planet formation scenarios. *Astrophys. J.* **745**, 174–188 (2012).
- Skrutskie, M. F. *et al.* The two micron all sky survey (2MASS). *Astron. J.* **131**, 1163–1183 (2006).
- Rebull, L. M. *et al.* The Taurus Spitzer survey: new candidate Taurus members selected using sensitive mid-infrared photometry. *Astrophys. J.* **186**, 259–307 (2010).
- Zhou, Y., Herczeg, G. J., Kraus, A. L., Metchev, S. & Cruz, K. L. Accretion onto planetary mass companions of low-mass young stars. *Astrophys. J.* **783**, L17–L22 (2014).
- Rigliaco, E. *et al.* X-shooter spectroscopy of young stellar objects. I. Mass accretion rates of low-mass T Tauri stars in  $\alpha$  Orionis. *J. Astron. Astrophys.* **548**, A56 (2012).
- Hartmann, L., Hewett, R. & Calvet, N. Magnetospheric accretion models for T Tauri stars. 1: Balmer line profiles without rotation. *Astrophys. J.* **426**, 669–687 (1994).
- Kenyon, S. J. & Hartmann, L. Pre-main-sequence evolution in the Taurus-Auriga molecular cloud. *Astrophys. J.* **101**, 117–171 (1995).
- Gullbring, E., Hartmann, L., Briceño, C. & Calvet, N. Disk accretion rates for T Tauri stars. *Astrophys. J.* **492**, 323–341 (1998).
- Ayliffe, B. A. & Bate, M. R. Migration of protoplanets with surfaces through discs with steep temperature gradients. *Mon. Not. R. Astron. Soc.* **415**, 576–586 (2011).
- Thalmann, C. *et al.* Optical imaging polarimetry of the LkCa 15 protoplanetary disk with SPHERE ZIMPOL. *Astrophys. J.* **808**, L41–L47 (2015).
- Isella, A., Chandler, C. J., Carpenter, J. M., Pérez, L. M. & Ricci, L. Searching for circumplanetary disks around LkCa 15. *Astrophys. J.* **788**, 129–135 (2014).

**Acknowledgements** This work was supported by NSF AAG grant no. 1211329 and NASA OSS grant NNX14AD20G. This material is based upon work supported by the National Science Foundation under grant no. 1228509. This work was performed in part under contract with the California Institute of Technology (Caltech) funded by NASA through the Sagan Fellowship Program executed by the NASA Exoplanet Science Institute. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under grant no. DGE-1143953. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors(s) and do not necessarily reflect the views of the National Science Foundation.

**Author Contributions** This work merged two independently acquired and analysed data sets. S.S. led preparation of the manuscript, the orbital fits, and the acquisition and analysis of the LBT data while K.B.F. led the acquisition and analysis of the MagAO data, development of the MagAO SDI pipeline, and drafted MagAO manuscript sections. S.S., K.B.F., J.E., L.C., P.H., A.S., J.M., and K.M. contributed to one or both observing proposals. J.E. modelled circumplanetary disk and hot-start scenarios, developed the NRM mode at LBT, and supervised effort of S.S.; L.C. carried out H $\alpha$  luminosity calculations and oversaw the MagAO effort. P.H. led LBTI development and support, and helped commission the NRM mode at LBT. K.K. carried out orbital stability analysis. J.M. developed the KLIP code used in MagAO data analysis. P.T. helped develop the NRM mode at LBT. B.M. supervised the effort of K.B.F.; S.S., K.B.F., J.E., L.C., and K.K. contributed key aspects of the manuscript. A.S., V.B., D.D., E.S., and A.V. supported the LBT observations. J.M., K.M., T.R., and A.W. supported the MagAO observations.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.S. (ssallum@email.arizona.edu).

## METHODS

**LBT observations and data reduction.** We observed LkCa 15 using non-redundant masking (NRM)<sup>11</sup> with LBTI/LMIRCam<sup>30,31</sup> in December 2014 and February 2015. NRM transforms a conventional telescope into an interferometric array through the use of a pupil-plane mask, providing better point-spread function (PSF) characterization and resolving angular scales even within  $\lambda/D$ . We used LMIRCam's 12-hole mask in single-aperture mode, yielding 1.4–7.0 m baselines. We broke up the observations into “visits,” each consisting of an identical set of integrations on LkCa 15 and an unresolved calibrator star (see Extended Data Table 1). We used three calibrators to lessen the possibility of contamination by a binary calibrator, and included one calibrator, GM Aur, from those observed previously at Keck<sup>8</sup>. We let the sky rotate throughout the observations, facilitating calibration of quasi-static speckles. At Ks and L' we observed LkCa 15 at parallactic angles between  $-37^\circ$  and  $65^\circ$ , and  $-65^\circ$  and  $65^\circ$ , respectively.

The NRM images show the interference fringes formed by the mask, the Fourier transform of which yields complex visibilities. Sampling the complex visibilities, we calculated squared visibilities (power versus baseline) and closure phases (sums of phases around three baselines forming a triangle). Closure phases eliminate atmospheric phase errors, leaving behind the sum of the intrinsic source phases. The LBT raw closure phase scatter was  $\sim 3^\circ$ , significantly lower than the uncertainties in previous NRM observations<sup>8</sup> ( $\sim 4^\circ$ ).

For each closing triangle, we fitted a polynomial to all calibrator closure phases as a function of time. We sampled the polynomial at the time of each target observation and subtracted it from each target closure phase. We calibrated using a variety of functions; of these, polynomials up to 2nd order in time provided the lowest-scatter calibrated closure phases, with standard deviations of  $1.7^\circ$  at Ks and  $1.9^\circ$  at L'. We calibrated the squared visibilities similarly, dividing by the calibrator rather than subtracting. We calibrated the mask baselines using the observed power spectra and knowledge of the filter bandpass and plate scale<sup>32</sup>.

**LBT image reconstruction, model fitting, and parameter error estimation.** We fitted models directly to kernel phases<sup>33,34</sup>, linearly independent combinations of closure phases, to search for companions. We modelled the central star as a delta function and each companion as another delta function located a distance  $s$  from the star, at position angle PA, and with contrast  $\Delta$ . We sampled the synthetic complex visibilities at the same baselines and parallactic angles as the data, and performed a grid fit, using a  $\Delta\chi^2$  to determine our parameter confidence intervals. Due to a known degeneracy between companion separation and contrast<sup>35</sup>, brighter companions at smaller separations provide equally good fits as those fainter and farther out. We thus performed fits to individual wavelengths to verify that the positions of b and c agreed across wavelength, then calculated a best fit where the companions coincided at Ks and L' (see Table 1). The model grids in this study required  $\sim 50,000$  CPU hours to generate, but were computed in a reasonable amount of time using the University of Arizona's El Gato supercomputer.

We also reconstructed images from the closure phases. To produce cleaner images, we re-calibrated the closure phases towards the best-fit Ks + L' model using an optimized calibrator weighting technique applied in previous NRM studies<sup>8</sup>. This calibration is similar to the locally optimized calibration of images (LOCI)<sup>36</sup> technique applied in direct imaging. Since this scheme can remove signal and underestimate errors, we applied it only to produce images (see Extended Data Fig. 1), using the polynomial calibration to estimate companion parameters. As a consistency check, we reconstructed images using both the BiSpectrum Maximum Entropy Method (BSMEM)<sup>37</sup> and the Monte-Carlo Markov Chain Imager algorithm (MACIM)<sup>38</sup>. The companion positions agree between the two algorithms, although BSMEM produces more compact emission towards each component. BSMEM has been shown to provide the most faithful image reconstruction of any available algorithms in blind tests<sup>39</sup>.

**Companion parameter error estimation for previously-published Keck data.** Orbital fitting required parameter errors for the previously published<sup>8</sup> Keck observations and the LBT observations to be consistently estimated. The published errors for the 2009–2010 companion parameters were generated using the nonlinear algorithm MPFIT<sup>40</sup>. While nonlinear fitters are often employed for computational expediency, the Levenberg–Marquardt algorithm can easily fall into a local minimum and underestimate parameter errors. The LBT grid  $\chi^2$  surfaces show local minima for both two- and three-companion fits, rendering MPFIT unreliable unless the starting parameters were very close to the global minimum. We compared MPFIT and grid-based parameter errors for the LBT data, and found that MPFIT significantly underestimated the errors (Fig. 2).

To create a “typical” error bar for each Keck companion, we estimated the error bar dependence on contrast using the LBT fits. Errors increased with decreasing companion flux, which we parameterized as a square root dependence. For a given Keck companion we thus scaled our LBT errors by the square root of the LBT-to-Keck flux ratio. We inflated the Keck error bars by a factor of 1.3, the ratio of the

uncalibrated closure phase scatter in the Keck data ( $\sim 4^\circ$ ) to that for the LBT data ( $\sim 3^\circ$ ). We capped the separation upper limits at  $3\lambda/D$ , where  $D$  is Keck's telescope diameter, 10 m, since the largest LBT upper limit was at nearly  $3\lambda/D$ , and companions at those distances are no longer subject to the separation-contrast degeneracy. **MagAO data reduction and analysis.** We observed LkCa 15 on November 15 and 22, 2014, as part of the Giant Accreting Protoplanet Survey (GAPplanetS), a visible-wavelength survey of bright transition disks. GAPplanetS stars are imaged simultaneously in H $\alpha$  ( $0.656\ \mu\text{m}$ ,  $\Delta\lambda = 6\ \text{nm}$ ) and the nearby stellar continuum ( $0.642\ \mu\text{m}$ ,  $\Delta\lambda = 6\ \text{nm}$ ) with the 585-actuator Magellan Adaptive Optics systems SDI camera<sup>12,41,42</sup>. The continuum channel provides a sensitive, simultaneous probe of the stellar PSF, allowing for effective removal of residual starlight and isolation of H $\alpha$  emitting sources<sup>12,43</sup>. The observations used new single-substrate narrowband H $\alpha$  and continuum filters, a significant improvement over the previous VisAO SDI filters, which suffered from ghost images<sup>12</sup>.

Seeing during the November 15 observations was better than the site median ( $0.56 \pm 0.06''$ ), winds were low ( $3.6 \pm 0.9\ \text{mph}$ ), and humidity was typical of the season ( $37.0 \pm 2.8\%$ ). Strehl ratio was low ( $< 10\%$ ), and difficult to measure meaningfully. We characterized image quality using the stellar full-width at half-maximum (FWHM),  $0.07''$  (at  $0.65\ \mu\text{m}$  over 30 s integrations), a significant improvement over the seeing. We collected 316 30-s closed-AO-loop images, with a total of 156 min of integration time and  $48.6^\circ$  of sky rotation. We selected the 149 LkCa 15 images with the lowest residual wavefront error ( $\sim 50\%$ ), equivalent to 74.5 min of exposure time. This image subset had  $47.6^\circ$  of sky rotation, with the rotational space well sampled.

The November 22 data were not of sufficient quality to recover LkCa 15 b, due to lower sky rotation ( $27.0^\circ$ ), shorter total integration (91 min), and shallower individual exposures (15 s). Injected positive planets with the same separation as LkCa 15 b were only recoverable with  $\text{SNR} > 3$  at contrasts  $> 5 \times 10^{-2}$  (nearly an order of magnitude brighter than the measured November 15 LkCa 15 b contrast). For this reason, we discuss only the November 15 data set in the rest of the paper.

Images were first bias-subtracted, registered, and aligned via cross-correlation. The H $\alpha$  flat field image showed very little non-uniformity across the field ( $< 1\%$ ), so a flat field was not applied. We masked CCD dust spots visible in the flat field wherever they affected the image throughput by more than 2%.

We processed the aligned data using angular differential imaging (ADI<sup>44</sup>), comparing the “classical” method of using a single median PSF for all images (cADI<sup>45</sup>) to the Karhunen–Loeve image processing (KLIP<sup>46</sup>) algorithm, which calculates a least-squares optimum PSF for each image. LkCa 15 b was detected in the H $\alpha$  channel via both methods, as shown in Extended Data Fig. 2. The planet was not detected in continuum with either method, so continuum images were used as a probe of PSF residuals and scattered light emission from the inner disk. Subtraction of the processed continuum images from the H $\alpha$  images (“ASDI”) left behind only true H $\alpha$  emission<sup>12</sup>.

**MagAO cADI reductions.** We constructed the stellar PSF by median combining images in  $0.5^\circ$  rotational bins and then median combining again to produce a PSF evenly sampled in rotational space. We subtracted the stellar PSF from the individual images, rotated them to a common on-sky orientation and combined them. Given the small separation between LkCa 15 A and b, the planet moved by only 1.5 FWHM over the course of the observations, resulting in self-subtraction and decreasing the FWHM of the processed planet PSF to 4–5 pixels in azimuth.

**MagAO KLIP-ADI reductions.** KLIP reductions were carried out using a well-tested custom IDL code<sup>47</sup>. To optimize reduction parameters, we maximized the signal to noise of injected planets (with the same separation and contrast as LkCa 15 b) inserted after using a negative planet to erase the LkCa 15 b signal. Planets were placed at position angles distant from known artefacts, and east or west of the star to avoid the noisier north/south region of the PSF, corresponding to the wind direction during the observations.

To limit self-subtraction, the library from which KLIP builds the stellar PSF is limited to images where a planet would have rotated away from its original position. We explored the size of this exclusion region (“rotational mask”) systematically through fake planet injection, and found that a  $5^\circ$  mask ( $\sim 1$  pixel at  $r = 11$  pixels) produced the highest signal-to-noise recoveries of injected planets. Given the stellar FWHM of 0.07, this resulted in azimuthal self-subtraction, with a processed planetary PSF of 2 pixels in azimuth.

Noise in the KLIP processed images was mostly Gaussian when images were divided into several independently-optimized radial zones, indicating efficient removal of speckles. Dividing these zones azimuthally provided no additional advantage, and the final KLIP reductions shown in Extended Data Fig. 2 reflect a PSF divided into 50-pixel ( $0.4''$ ) annuli. Removal of the median PSF radial profile for the entire image set aided significantly in attenuating the stellar halo, improving the ability of the KLIP algorithm to match residual speckles and enhancing contrast close to the star.



**MagAO LkCa 15 b photometry and astrometry.** We estimated photometry and astrometry by minimizing residuals after injecting a negative planet at the location of LkCa 15 b. The cube of registered and aligned H $\alpha$  channel images was scaled by the chosen contrast value, multiplied by  $-1$ , and injected into the raw images before KLIP processing. Using the full H $\alpha$  image cube rather than its median combination simulated variability of the PSF between images.

We generated error bars by injecting false positive planets with similar separations and contrasts to LkCa 15 b after using a negative planet to eliminate the true signal. Planets were placed at position angles away from the wind direction, and spaced by at least 75% of the measured stellar FWHM. We computed the centroid and peak pixel using a 5-pixel aperture around each planet, and assigned the standard deviations in recovered flux and position as our  $1\sigma$  photometric and astrometric uncertainties, respectively (see Extended Data Table 2 and Extended Data Fig. 3).

**Signal-to-noise of the MagAO H $\alpha$  detection.** To create signal-to-noise ratio (SNR) maps, we calculated a radial noise profile using the standard deviation of 1-pixel-wide annuli and divided it into the raw images. In the raw maps, LkCa 15 b has  $\text{SNR} \approx 3$ –4. Smoothing by a Gaussian with a 2-pixel FWHM maximized the SNR of injected fake planets, so we applied this smoothing to the final science images, resulting in peak SNRs of 4.4 and 6.8 in the KLIP H $\alpha$  and ASDI images, respectively. However, directly-imaged exoplanets at small separations suffer from small number statistical effects<sup>48</sup>. The underlying speckle distribution is difficult to probe given the small number of independently sampled noise regions. In an annulus at the distance of LkCa 15 b (1.3 FWHM), seven noise regions exist, leading to corrected<sup>48</sup> SNRs of 4.1 and 6.4 for the H $\alpha$  and ASDI images, respectively. The ASDI detection corresponds to a false positive probability of  $3 \times 10^{-4}$  using the Student's  $t$ -distribution with 6 degrees of freedom.

Comparing the LkCa 15 b SNR to the distribution of values in the ASDI SNR map (Extended Data Fig. 4), shows that it is a clear outlier. Comparison of the peak pixel in an aperture centred on b to those in the surrounding noise apertures (Extended Data Figs 4, 5) further demonstrates b's statistical significance.

In addition to the high SNR, low false positive fraction, and the statistics presented in Extended Data Fig. 4, the H $\alpha$  detection is significant because it occurs at the same location as the independent LBT detection. This further reduces the probability of a false positive detection in the MagAO data, since speckles have no preferred location.

**Fidelity of the MagAO LkCa 15 b detection.** Neither the existence of LkCa 15 b nor its derived parameters are dependent on our choice to include only the top 50% of raw images. The planet appears at the same location and with the same approximate brightness when processing all 316 images, as well as only the top 25% of images. An excess with  $\text{SNR} > 3$  appears at LkCa 15 b's location with a wide range of KLIP zone geometries and rotational masks, when any number of KL modes from 2 to 100+ are removed, and whether or not the median radial profile of the PSF is subtracted before processing.

**Limits on MagAO LkCa 15 b continuum flux.** We used simulated planet detections to place an upper limit on LkCa 15 b's continuum flux. We injected planets into the raw continuum channel images with a range of contrasts and at positions near LkCa 15 b. We then measured the SNR of each simulated detection to determine the confidence at which we could detect a given contrast. As above, we apply a small number statistical correction<sup>48</sup> to the SNR of each recovered planet. The simulations suggest that we would have detected an excess with a corrected SNR of 3 (false positive fraction of  $10^{-2}$ ) for a continuum source with contrast greater than  $5 \times 10^{-3}$ . Since LkCa 15 A is 1.8 times brighter at H $\alpha$  than continuum, this corresponds to an H $\alpha$ -to-continuum-flux ratio lower limit of 2.7.

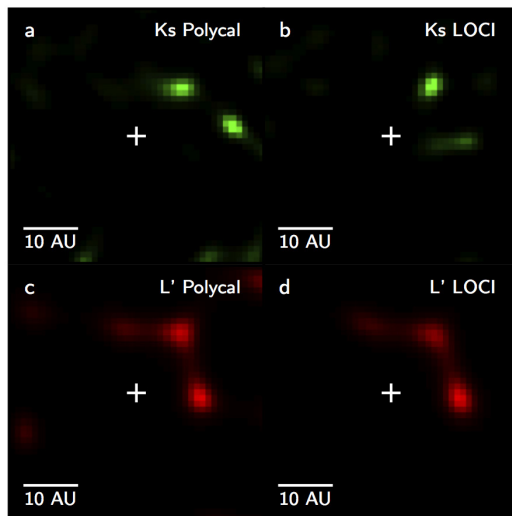
**Limits on MagAO LkCa 15 c H $\alpha$  contrast.** We established limits on the LkCa 15 c H $\alpha$  contrast using false planet injections, first using a negative planet to eliminate the LkCa 15 b signal. We injected planets with a range of contrasts at positions sampling the LBT error ellipse for LkCa 15 c. At position angles between  $-40^\circ$  and  $-52^\circ$ , several  $2$ – $2.5\sigma$  peaks near c's location boost the SNRs for recovered planets. Here, we can detect contrasts down to  $2 \times 10^{-3}$  with corrected<sup>48</sup> SNRs of 3 (false positive fraction of  $2 \times 10^{-2}$ ). Position angles greater than  $-40^\circ$  approach a noisier region of the PSF, leading to decreased sensitivity; here contrasts of  $6 \times 10^{-3}$  are required for adjusted signal-to-noise ratios of 3. We cannot reject or confirm accretion onto LkCa 15 c below  $6 \times 10^{-3}$  contrast ( $\Delta H\alpha = 5.6$  mag) with the current data set. This improves upon previous spectro-astrometric observations, which yielded a contrast limit of  $\Delta H\alpha = 3.4$  mag for a position angle near LkCa 15 c<sup>10</sup>.

**Stability analysis with LkCa 15 d.** We ran a series of orbit integrations to demonstrate that stable solutions exist for b, c, and d at separations within the  $1\sigma$  semimajor axis error bars (see Extended Data Figs 6, 7). We used the publicly

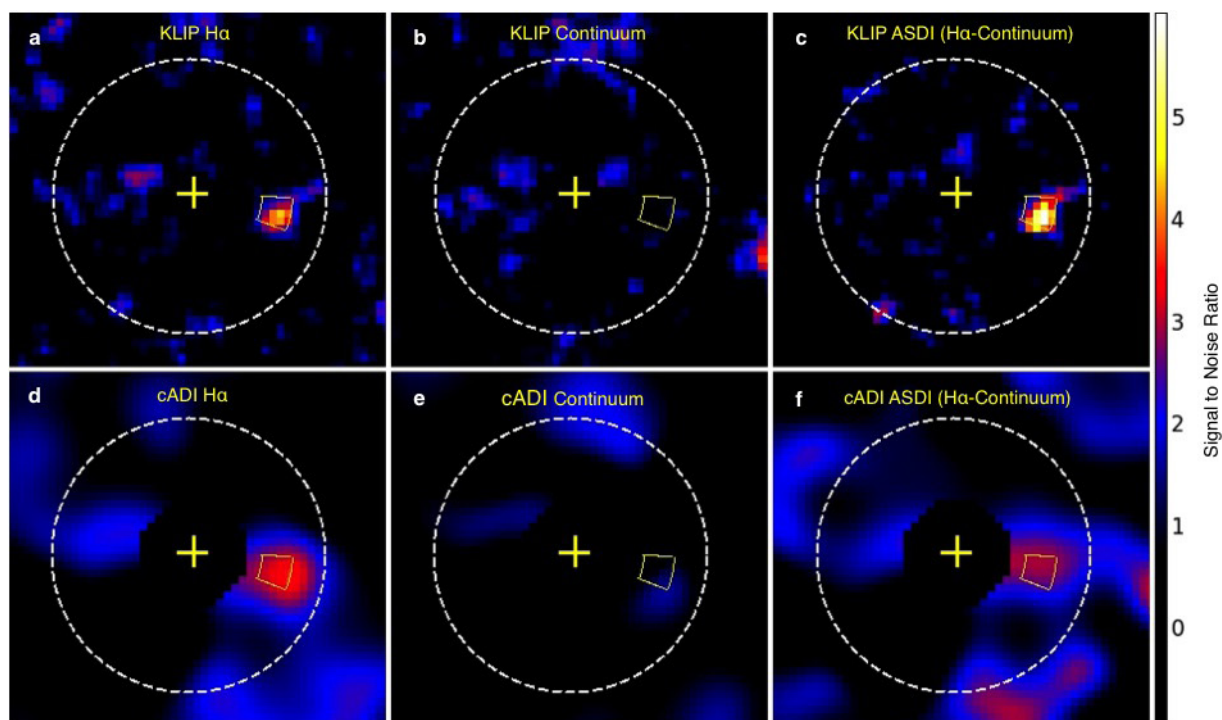
available Swifter package<sup>49</sup>—specifically, the symplectic integrator, SyMBA<sup>50</sup>, which switches to a Burlisch-Stoer algorithm for planetary close approaches. We also ran comparison integrations with the Gauss Radau 15th order integrator and found comparable results, with minimum energy conservation of 1 part in  $10^7$  over a 10 Myr integration.

We required all orbits to be nearly co-planar, with a random scatter  $< 1^\circ$ , and assigned each planet a random eccentricity below  $10^{-4}$ . To assess stability we integrated three different random phase combinations for 10 Myr. We found stable three body solutions out to 1–2 Myr with semi-major axes of  $a_b = 12.7$  AU,  $a_c = 18.6$  AU,  $a_d = 24.7$  AU. To ensure stability out to 10 Myr with orbits in the  $1\sigma$  errors requires that all planets be  $\leq 0.5 M_J$ . A wider range of orbits is allowed if d's mass is decreased further. These constraints are in line with previous large numerical studies of equally spaced (in  $R_{H,m}$ ), equal-mass planets<sup>51</sup>. Note for planets b and c, there are possible resonant configurations within the predicted period ranges, which would admit somewhat higher masses.

30. Hinz, P. M. *et al.* Status of the LBT interferometer. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series* Vol. 7013 28–36 (2008).
31. Leisenring, J. M. *et al.* On-sky operations and performance of LMIrcam at the Large Binocular Telescope. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series* Vol. 8446 4–19 (2012).
32. Maire, A.-L. *et al.* The LEECH exoplanet imaging survey. Further constraints on the planet architecture of the HR 8799 system. *J. Astron. Astrophys.* **576**, A133 (2015).
33. Martinache, F. Kernel phase in Fizeau interferometry. *Astrophys. J.* **724**, 464–469 (2010).
34. Ireland, M. J. Phase errors in diffraction-limited imaging: contrast limits for sparse aperture masking. *Mon. Not. R. Astron. Soc.* **433**, 1718–1728 (2013).
35. Sallum, S. *et al.* New spatially resolved observations of the T Cha transition disk and constraints on the previously claimed substellar companion. *Astrophys. J.* **801**, 85–107 (2015).
36. Lafrenière, D., Marois, C., Doyon, R., Nadeau, D. & Artigau, É. A new algorithm for point-spread function subtraction in high-contrast imaging: a demonstration with angular differential imaging. *Astrophys. J.* **660**, 770–780 (2007).
37. Buscher, D. F. Direct maximum-entropy image reconstruction from the bispectrum. In *Very High Angular Resolution Imaging* (eds Robertson, J. G. & Tango, W. J.) 91–93 (Vol. 158 of IAU Symposium, 1994).
38. Ireland, M. J., Monnier, J. D. & Thureau, N. Monte-Carlo imaging for optical interferometry. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series* Vol. 6268 1T1–1T8 (2006).
39. Lawson, P. R. *et al.* 2006 interferometry imaging beauty contest. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series* Vol. 6268 1U1–1U12 (2006).
40. Markwardt, C. B. Non-linear Least-squares Fitting in IDL with MPFIT. In *Astronomical Data Analysis Software and Systems XVIII* (eds Bohlender, D. A., Durand, D. & Dowler, P.) Vol. 411 of *Astronomical Society of the Pacific Conference Series*, 251 (2009).
41. Morzinski, K. *et al.* MagAO: Status and on-sky performance of the Magellan adaptive optics system. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series* Vol. 914804 1–13 (2014).
42. Close, L. *et al.* First closed-loop visible AO test results for the advanced adaptive secondary AO system for the Magellan Telescope: MagAO's performance and status. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series* Vol. 8447 0X1–0X16 (2012).
43. Follette, K. B. *et al.* The first circumstellar disk imaged in silhouette at visible wavelengths with adaptive optics: MagAO imaging of Orion 218–354. *Astrophys. J.* **775**, L13–L17 (2013).
44. Marois, C. *et al.* Direct imaging of multiple planets orbiting the star HR 8799. *Science* **322**, 1348–1352 (2008).
45. Marois, C., Lafrenière, D., Doyon, R., Macintosh, B. & Nadeau, D. Angular differential imaging: a powerful high-contrast imaging technique. *Astrophys. J.* **641**, 556–564 (2006).
46. Soummer, R., Pueyo, L. & Larkin, J. Detection and characterization of exoplanets and disks using projections on Karhunen-Loève eigenimages. *Astrophys. J.* **755**, L28–L32 (2012).
47. Males, J. R. *et al.* Magellan adaptive optics first-light observations of the exoplanet  $\beta$  Pic b. I. Direct imaging in the far-red optical with MagAO+VisAO and in the near-IR with NICI. *Astrophys. J.* **786**, 32–53 (2014).
48. Mawet, D. *et al.* Fundamental limitations of high contrast imaging set by small sample statistics. *Astrophys. J.* **792**, 97–107 (2014).
49. Levison, H. F. & Duncan, M. J. SWIFT: A solar system integration software package. *Astrophysics Source Code Library* (2013).
50. Duncan, M. J., Levison, H. F. & Lee, M. H. A multiple time step symplectic algorithm for integrating close encounters. *Astron. J.* **116**, 2067–2077 (1998).
51. Faber, P. & Quillen, A. C. The total number of giant planets in debris discs with central clearings. *Mon. Not. R. Astron. Soc.* **382**, 1823–1828 (2007).



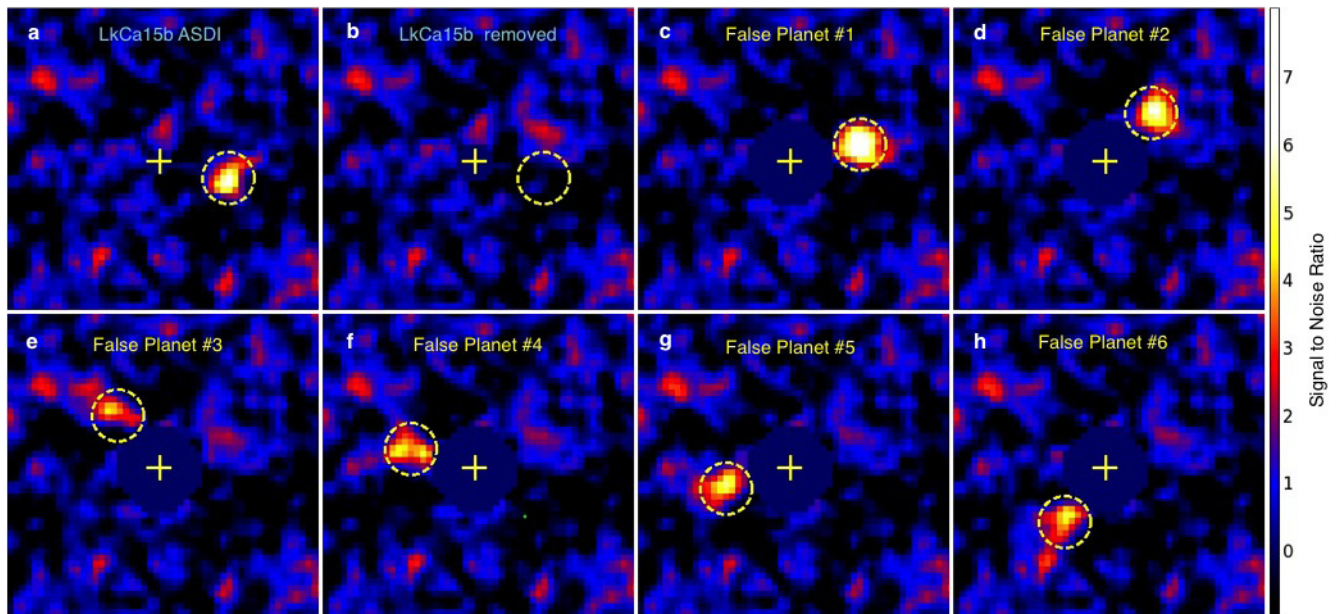
**Extended Data Figure 1 | Image reconstructions.** **a–d**, Images reconstructed from closure phases, showing Ks polynomial (**a**) and LOCI-like (**b**) calibrations, and L' polynomial (**c**) and LOCI-like (**d**) calibrations. Both calibrations yielded reconstructed images with at least two distinct components. The LOCI-like calibration moved each companion within the position errors derived from the grid  $\chi^2$  surface.



**Extended Data Figure 2 | KLIP and ADI H $\alpha$  SNR maps.** **a–c**, Final KLIP SNR maps for H $\alpha$  (**a**), continuum (**b**) and the difference between the two (ASDI, **c**). **d–f**, Final cADI SNR maps in the same order. Dividing by the radial noise profiles to create these maps should normalize the noise distribution at all radii within the speckle-dominated regime.

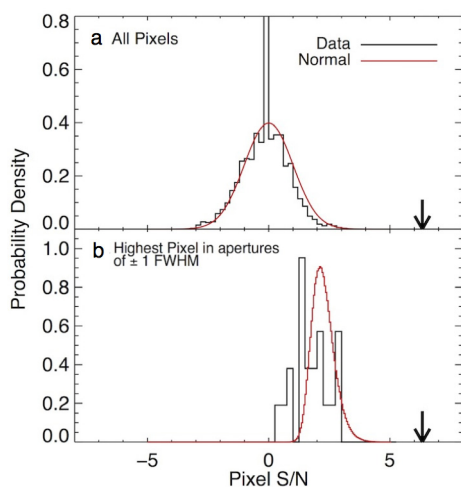
The presence of dark holes in the maps suggests that we are speckle-dominated out to the AO control radius at  $r \approx 20$  pixels (white, dashed circles). LkCa 15 b's separation is 11.6 pixels. The yellow keystones indicate the  $2\sigma$  range of allowed astrometry for the KLIP ASDI point source (upper right) based on negative simulated planet injection.



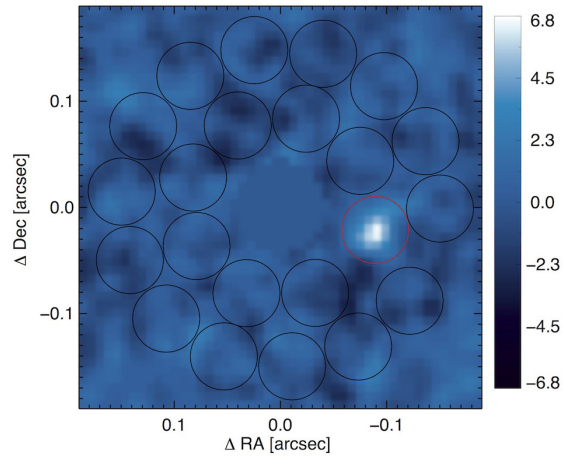


**Extended Data Figure 3 | False positive planet SNR maps.** **a**, LkCa 15 final ASDI SNR map. **b**, ASDI SNR map with LkCa 15 b removed. **c–h**, ASDI SNR maps of false positive planets injected at a radius of

11 pixels and contrast of  $8 \times 10^{-3}$ . Recovered parameters for these planets are given in Extended Data Table 2 and were used to determine  $1\sigma$  astrometric and photometric uncertainties.

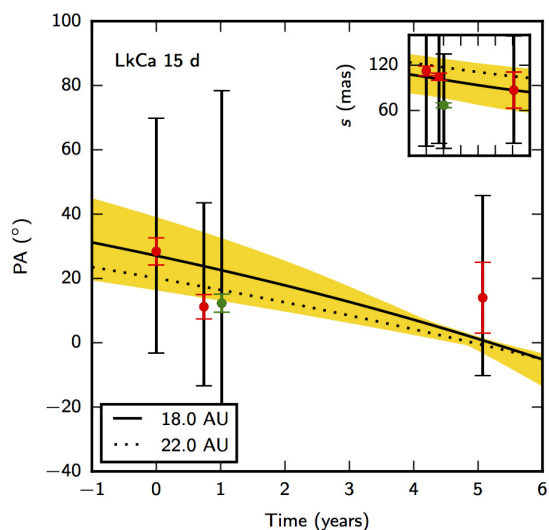


**Extended Data Figure 4 | H $\alpha$  detection noise statistics.** **a**, Histogram of noise (non-planet) pixel values in the SNR map within the speckle dominated regime (black line) compared to a Normal distribution (red line). The black arrow denotes the location of the peak SNR value for LkCa 15 b. **b**, Histogram of the peak values in all noise apertures (see Extended Data Fig. 5) within the control radius (black line) compared to a Normal distribution (red line). The black arrow shows the peak pixel value in the LkCa 15 b aperture.

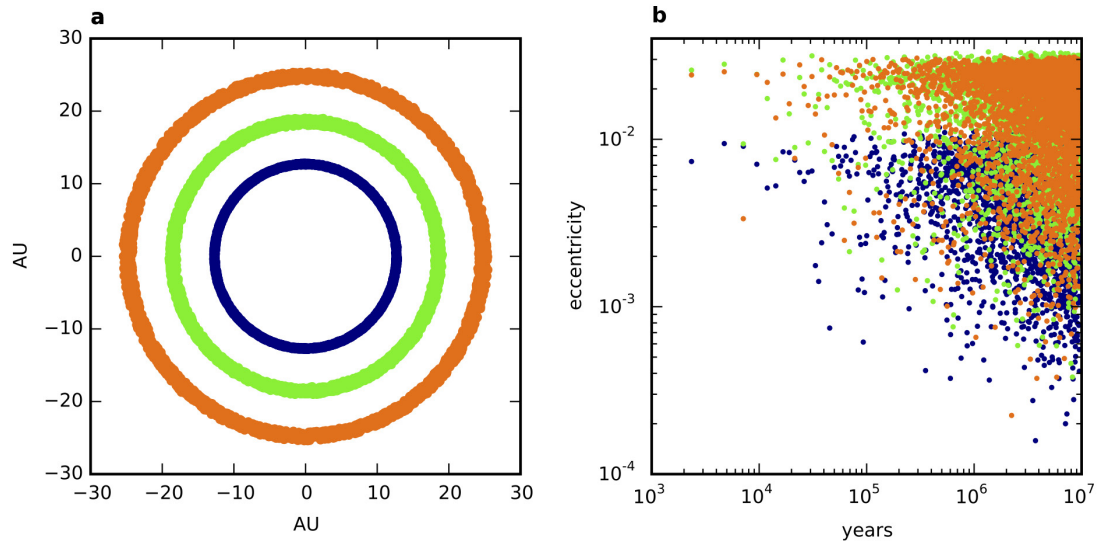


**Extended Data Figure 5 | Noise apertures.** Noise apertures (black circles) surrounding LkCa 15 A used to calculate the statistics presented in Extended Data Fig. 4. Colour indicates SNR.





**Extended Data Figure 6 | LkCa 15 d position angle and separation versus time.** Evolution of position angle and separation (inset) for LkCa 15 d. Green and red points indicate Ks and L' data, respectively. In both panels, the earliest three points correspond to previously published Keck observations<sup>8</sup>, and the most recent points show best fits to our data. The coloured error bars are derived using the nonlinear algorithm MPFIT, which significantly underestimates the parameter errors compared to the more robust grid,  $\Delta\chi^2$  (black error bars; see Methods). The yellow shaded region spans the position angles and separations allowed at  $1\sigma$  by the multi-epoch observations, which have semi-major axes between 12.6 and 24.7 AU. Solid curves show the best-fit orbit (18.0 AU), and dashed curves show an orbit (24.7 AU) that is stable for a  $0.5 M_J$  planet exterior to LkCa 15 b and c. Lower mass planets or resonant configurations permit stable orbits for LkCa 15 d at smaller stellocentric radii.



**Extended Data Figure 7 | Orbital integration results.** **a**, Stable orbits for LkCa 15 b, c, and d over a 10 Myr integration. **b**, Osculating eccentricity. The planets are each  $0.5 M_J$  with initial semi-major axes of 12.7, 18.6, and

24.7 AU, initial eccentricities of order  $10^{-5}$ , and relative inclinations of  $< 1^\circ$ . After a 10 Myr integration, the eccentricities of c and d have increased to only a few percent.

Extended Data Table 1 | Summary of observations

Target	Right Ascension (hh mm ss.sss)	Declination (dd mm ss.sss)	t <sub>int</sub> (s)	N <sub>frames</sub> *	N <sub>visits</sub> †	Total Time (h)	Average Seeing (asec)
2014 Nov 15: H $\alpha$ and 642 nm continuum							
LkCa 15	04 39 17.796	+22 21 03.48	30	316	1	2.63	0.56
2014 Nov 22: H $\alpha$ and 642 nm continuum							
LkCa 15	04 39 17.796	+22 21 03.48	15	364	1	1.52	N/A‡
2014 Dec 15: L'							
LkCa 15	04 39 17.796	+22 21 03.48	10	40	15	1.67	0.76
HD284668	04 42 09.686	+22 13 55.62	10	40	5	0.56	
HD284581	04 40 32.495	+22 31 32.88	10	40	4	0.44	
GM Aur	04 55 10.983	+30 21 59.54	10	40	5	0.56	
2015 Feb 5-7: Ks							
LkCa 15	04 39 17.796	+22 21 03.48	20	20	19	2.11	0.95
HD284668	04 42 09.686	+22 13 55.62	20	20	7	0.78	
HD284581	04 40 32.495	+22 31 32.88	20	20	7	0.78	
GM Aur	04 55 10.983	+30 21 59.54	20	20	6	0.67	

\*Number of frames in each visit.

†Each visit consists of all images taken before switching between target and calibrator.

‡The seeing monitor was unavailable during the 22 November observations.



Extended Data Table 2 | False planet injection results

P.A. *	$X_{in}^{\dagger}$	$Y_{in}^{\ddagger}$	$X_{rec}^{\S}$	$Y_{rec}^{\parallel}$	$\Delta P.A.^{\P}$	$\Delta s^{\#}$	Peak SNR
( $^{\circ}$ )	(pix)	(pix)	(pix)	(pix)	( $^{\circ}$ )	(pix)	
-77	135.2	127.0	136.0	127.3	-0.53	-0.85	8.9
-44	132.1	132.4	132.8	132.8	1.11	-0.78	6.2
38	117.7	133.2	116.2	134.3	-2.25	-1.80	4.3
73	114.0	127.7	112.3	127.8	-1.81	-1.66	4.6
108	114.0	121.1	113.9	122.0	4.67	0.15	5.5
143	117.9	115.7	118.1	116.7	2.50	0.91	4.7
Simulated Planet Means					0.62	-0.67	5.7

## LkCa 15 b Fit Results

	X	Y	P.A.	s	Peak SNR
Parameters	135.8	121.8	-103.4	11.6	6.8
1 $\sigma$ Errors			$\pm 2.7$	$\pm 1.0$	$\pm 30\%$

\*Input false planet position angle.

 $\dagger$ Input false planet X location. $\ddagger$ Input false planet Y location. $\S$ Recovered false planet X location. $\parallel$ Recovered false planet Y location. $\P$ Recovered minus input false planet position angle. $\#$ Recovered minus input false planet separation.

# Measurement of interaction between antiprotons

The STAR Collaboration\*

One of the primary goals of nuclear physics is to understand the force between nucleons, which is a necessary step for understanding the structure of nuclei and how nuclei interact with each other. Rutherford discovered the atomic nucleus in 1911, and the large body of knowledge about the nuclear force that has since been acquired was derived from studies made on nucleons or nuclei. Although antinuclei up to antihelium-4 have been discovered<sup>1</sup> and their masses measured, little is known directly about the nuclear force between antinucleons. Here, we study antiproton pair correlations among data collected by the STAR experiment<sup>2</sup> at the Relativistic Heavy Ion Collider (RHIC)<sup>3</sup>, where gold ions are collided with a centre-of-mass energy of 200 gigaelectronvolts per nucleon pair. Antiprotons are abundantly produced in such collisions, thus making it feasible to study details of the antiproton–antiproton interaction. By applying a technique similar to Hanbury Brown and Twiss intensity interferometry<sup>4</sup>, we show that the force between two antiprotons is attractive. In addition, we report two key parameters that characterize the corresponding strong interaction: the scattering length and the effective range of the interaction. Our measured parameters are consistent within errors with the corresponding values for proton–proton interactions. Our results provide direct information on the interaction between two antiprotons, one of the simplest systems of antinucleons, and so are fundamental to understanding the structure of more-complex antinuclei and their properties.

Although the theory of quantum chromodynamics (QCD) provides us with an understanding of the foundation of the nuclear force, this binding interaction in nuclei operates at low energy, where the force is strong and difficult to calculate directly from the theory (see ref. 5 and references therein for recent developments). For that reason, a common parameterization of the effective interaction between nucleons is based on experimental measurements, and the corresponding parameterization for antinucleons remains undetermined. The important parameters in such a description of the interaction are the scattering length ( $f_0$ ), which is related to elastic cross-sections, and the effective range of the interaction ( $d_0$ ), which is determined to be a few femtometres (the typical nuclear scale). For a short range potential, these two parameters are related to the  $s$ -wave scattering phase shift  $\delta_0$  and the collision momentum  $k$  by  $k \cot(\delta_0) \approx \frac{1}{f_0} + \frac{1}{2}d_0k^2$ . The existence and

production rates of antinuclei offer indirect information about interactions between antinucleons, and also have relevance to the unexplained baryon asymmetry in the Universe<sup>6</sup>. Antinuclei produced to date include antiprotons, antideuteron, antitriton, antihelium-3, and the recently discovered antihypertriton and antihelium-4 (see ref. 1 and references therein). The interaction between two antinucleons is the basic interaction that binds the antinucleons into antinuclei, and this has not been directly measured previously. Of equal importance, one aspect of the current measurement is a test of matter–antimatter symmetry, more formally known as CPT—a fundamental symmetry of physical laws under the simultaneous transformations of charge conjugation (C), parity transformation (P) and time reversal (T). Although various prior CPT tests<sup>7</sup> have been many orders of

magnitude more precise than what is reported here, there is value in independently verifying each distinct prediction of CPT symmetry<sup>7</sup>.

Ultra-relativistic nuclear collisions produce an energy density similar to that of the Universe microseconds after the Big Bang, and the high energy density creates a favourable environment for antimatter production. The abundantly produced antiprotons provide the opportunity to measure, for the first time, the parameters  $f_0$  and  $d_0$  of the strong nuclear force between antinucleons rather than nucleons.

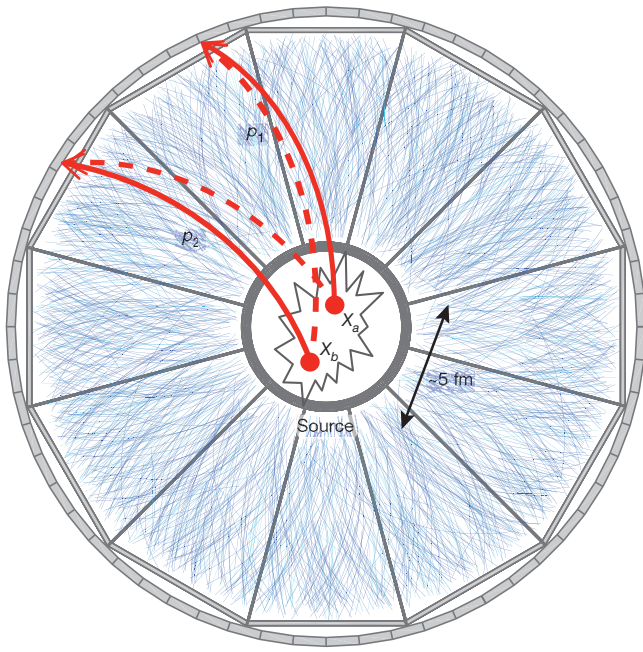
The technique used to probe the antiproton–antiproton interaction involves momentum correlations, and it resembles the space-time correlation technique used in HBT (Hanbury Brown and Twiss) intensity interferometry. Since its invention for use in astronomy in the 1950s<sup>4</sup>, the HBT technique has been adopted in many areas of physics, including the study of the quantum state of Bose–Einstein condensates<sup>8</sup>, and the correlation among electrons<sup>9</sup> and among atoms in cold Fermi gases<sup>10</sup>. A Bose–Einstein enhancement in particle physics was first observed in the late 1950s as an enhanced number of pairs of identical pions produced with small opening angles, the GGLP (Goldhaber, Goldhaber, Lee and Pais) effect<sup>11</sup>. Later on, Kopylov and Podgoretsky noted the common quantum statistics origin of the HBT and GGLP effects<sup>12</sup>, and, through a series of papers (see a review<sup>13</sup> and references therein), they devised the basics of the momentum correlation interferometry technique. In this technique, they introduced the correlation functions (CFs) as ratios of the momentum distributions of correlated and uncorrelated particles,  $C(\mathbf{p}_1, \mathbf{p}_2) = \frac{P(\mathbf{p}_1, \mathbf{p}_2)}{P(\mathbf{p}_1)P(\mathbf{p}_2)}$  with

$C = 1$  for no correlations, suggested the so-called mixing technique to construct the uncorrelated distribution by using particles from different collisions (events), and formulated a simple relation of the CFs with the space-time structure of the particle emission region. Here  $C(\mathbf{p}_1, \mathbf{p}_2)$  is the correlation function,  $P(\mathbf{p}_1)$  and  $P(\mathbf{p}_2)$  are probabilities for detecting a particle with momentum  $\mathbf{p}_1$  and a particle with momentum  $\mathbf{p}_2$ , respectively, and  $P(\mathbf{p}_1, \mathbf{p}_2)$  is the joint probability for detecting both simultaneously. As a result, the momentum correlation technique has been widely embraced by the nuclear physics community<sup>14–17</sup>.

Figure 1 illustrates the process of constructing two-particle correlations in heavy-ion collisions. In addition to quantum statistics effects, final state interactions (FSIs) play an important role in the formation of correlations between particles. FSIs include, but are not limited to, the formation of resonances, the Coulomb repulsion effect, and the nuclear interactions between two particles<sup>14,15,18,19</sup>. In fact, FSI effects provide valuable additional information. They allow for (see refs 16, 20 and references therein) coalescence femtoscopy, correlation femtoscopy with non-identical particles, including access to the relative space-time production asymmetries, and a measurement of the strong interaction between specific particles. The last measurement is often difficult to access by other means and is the focus of this paper (for recent studies see refs 21, 22).

In a semi-classical geometrical description, a complex heavy-ion collision can be regarded as a superposition of many individual nucleon–nucleon collisions, each governed by a constant probability of interaction with all nucleons travelling in straight lines. The centrality corresponds to the extent that two nuclei overlap, and events are categorized by their

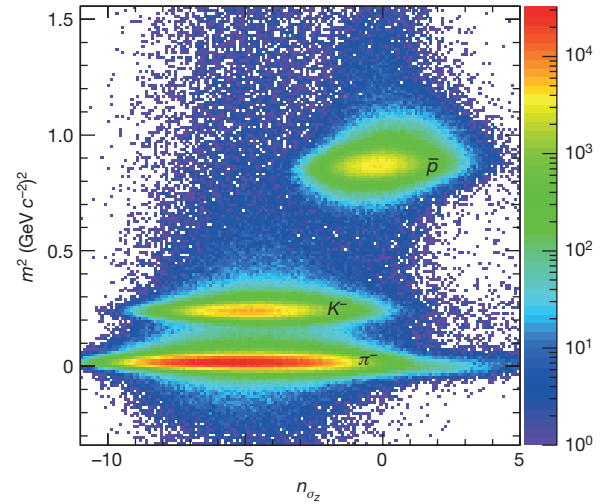
\*Lists of participants and their affiliations appear at the end of the paper.



**Figure 1 | A schematic of the two-particle correlation process in a heavy-ion collision.** The display is overlaid on an event display from the Time Projection Chamber in the STAR detector. The curves show particle trajectories, from which the track momenta are determined. These trajectories are measured in three dimensions, but are projected onto a single plane in this illustration. The STAR detector measures three-vector momenta over a wide range beginning at about  $0.1 \text{ GeV } c^{-1}$ . Two particles emitted from two separated points, with four-coordinates  $X_a$  and  $X_b$ , are detected with four-momenta  $p_1$  and  $p_2$ . For the pair of indistinguishable particles with even/odd total spin, the two quantum mechanical amplitudes (representing, for non-interacting particles, products of plane waves  $\langle p_1|X_a \rangle \langle p_2|X_b \rangle$  and  $\langle p_2|X_a \rangle \langle p_1|X_b \rangle$ , where  $\langle p|X \rangle = \exp(-ipX)$ ) must be added/subtracted to yield the amplitude which is symmetric/antisymmetric with respect to the interchange of particle momenta. This results in an enhancement/suppression in the joint detection probability at zero momentum separation with the width inversely proportional to the space-time separation of particle emission points.

centrality, based on the observed number of tracks emitted from each collision. Zero per cent centrality corresponds to exactly head-on collisions which produce the most tracks, while 100% centrality corresponds to barely glancing collisions which produce the fewest tracks. The data used here consists of Au + Au collisions at a centre-of-mass energy of 200 GeV per nucleon pair, taken during the operation of RHIC in the year 2011. In total, 500 million events were taken by the minimum-bias trigger at STAR. This trigger selects all particle-producing collisions, regardless of the extent of overlap of the incident nuclei, but with a requirement that collisions must have occurred along the trajectory of the colliding Au ion and within  $\pm 30 \text{ cm}$  of the centre of STAR's Time Projection Chamber (TPC)<sup>23</sup>. Events used in this analysis correspond to the 30%–80% centrality class, for which the signal due to two-particle interaction is stronger than that from smaller centrality classes, while particle yields are larger than that from larger centrality classes.

The two main detectors used in the measurement are the STAR TPC and the Time of Flight Barrel (TOF)<sup>24</sup>. The TPC is situated in a solenoidal magnetic field (0.5 T), and it provides a three-dimensional image of the ionization trails left along the path of charged particles. The TOF encloses the curved surface of the cylindrical TPC. In conjunction with the momentum measured via the track curvature in TPC, particle identification (PID) is achieved by two key measurements: the mean energy loss per unit track length,  $\langle dE/dx \rangle$ , which can be used to distinguish particles with different masses or charges, and the time of flight of particles reaching the TOF detector, which can be used, together with tracking information, to derive the square of a



**Figure 2 |  $m^2$  versus  $n_{\sigma_z}$  for negatively charged particles.** Here  $m^2 = (p^2/c^2)(t^2c^2/L^2 - 1)$ , where  $t$  and  $L$  are the time of flight and path length, respectively.  $c$  is the light velocity.  $z = \ln(\langle dE/dx \rangle / \langle dE/dx \rangle_E)$  and  $\langle dE/dx \rangle_E$  is the expected value of  $\langle dE/dx \rangle$  for (anti)protons.  $\sigma_z$  is the r.m.s. width of the  $z$  distribution, and  $n_{\sigma_z}$  is the number of standard deviations from zero, the expected value of  $z$  for (anti)protons. The antiprotons, centred at  $m^2 = 0.88 (\text{GeV } c^{-2})^2$  and  $n_{\sigma_z} = 0$ , are well separated from other particle species. (Anti)protons satisfying  $0.8 (\text{GeV } c^{-2})^2 < \text{mass}^2 < 1 (\text{GeV } c^{-2})^2$  and  $|n_{\sigma_z}| < 1.5$  are selected for making pairs. With this selection, the purity is  $> 99\%$  for (anti)protons with transverse momentum less than  $2 \text{ GeV } c^{-1}$ . Colours denote particle population (counts) in cells formed by even division of  $m^2$  and  $n_{\sigma_z}$ .

particle's mass ( $m^2$ ). Figure 2 shows a typical calculated mass-squared ( $m^2$ ) distribution versus  $n_{\sigma_z}$  (see Fig. 2 legend) for antiprotons.

The population distribution of (anti)proton pairs as a function of (anti)proton momentum ( $k^*$ ) in the pair rest frame (in which the centre of mass of the pair is at rest, convenient for carrying out measurements) is measured for the correlated pairs from within the same event,  $A(k^*)$ , and, separately, for the non-correlated pairs from two different (mixed) events,  $B(k^*)$ . The former corresponds to the joint probability  $P(p_1, p_2)$ , and the latter corresponds to the product of two probabilities,  $P(p_1)P(p_2)$ , where  $P(p_1)$  and  $P(p_2)$  each corresponds for observing single (anti)protons. The ratio of the two,  $A(k^*)/B(k^*)$ , gives the measured CF (see Methods). The observed (anti)protons can come from weak decays of already correlated primary particles, hence introducing residual correlations which contaminate the CF. The dominant contaminations to the CF come from the  $p$ - $\Lambda$  ( $\bar{p}$ - $\bar{\Lambda}$ ) and  $\Lambda$ - $\Lambda$  ( $\bar{\Lambda}$ - $\bar{\Lambda}$ ) correlations (where  $p$  and  $\Lambda$  denotes the proton and lambda particle, respectively, and  $\bar{p}$  and  $\bar{\Lambda}$  denotes the corresponding antiparticle), and are taken into account by fitting the CF with corresponding contributions. Taking the two-proton correlation measurement as an example<sup>25</sup>,

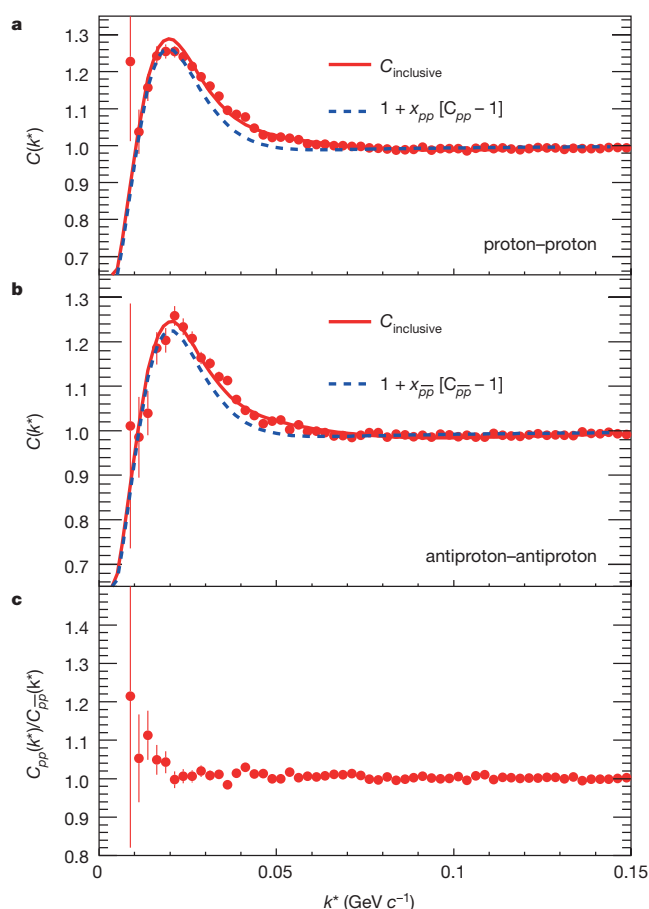
$$C_{\text{inclusive}}(k^*) = 1 + x_{pp}[C_{pp}(k^*; R_{pp}) - 1] + x_{p\Lambda}[\tilde{C}_{p\Lambda}(k^*; R_{p\Lambda}) - 1] + x_{\Lambda\Lambda}[\tilde{C}_{\Lambda\Lambda}(k^*) - 1] \quad (1)$$

where  $C_{\text{inclusive}}(k^*)$  is the inclusive CF, and  $C_{pp}(k^*; R_{pp})$  is the true proton–proton CF, which can be described by the Lednický and Lyuboshitz analytical model<sup>19</sup>. In this model, for given  $s$ -wave scattering parameters, the correlation function with FSI is calculated as the square of the properly symmetrized wavefunction averaged over the total pair spin and the distribution of relative distances of particle emission points in the pair rest frame (see Methods).  $\tilde{C}$  are the residual CFs which are expressed through the  $p$ - $\Lambda$  and  $\Lambda$ - $\Lambda$  CFs,  $C_{p\Lambda}(k_{p\Lambda}^*; R_{p\Lambda})$  and  $C_{\Lambda\Lambda}(k_{\Lambda\Lambda}^*)$ , using integral transformation<sup>25</sup> from  $k_{p\Lambda}^*$  and  $k_{\Lambda\Lambda}^*$  to  $k_{pp}^*$  (see Methods).  $C_{p\Lambda}(k_{p\Lambda}^*; R_{p\Lambda})$  is taken from a theoretical calculation<sup>19</sup>, which includes all final-state interactions and explains experimental data well<sup>21</sup>.  $C_{\Lambda\Lambda}(k_{\Lambda\Lambda}^*)$  is from an experimental measurement corrected for

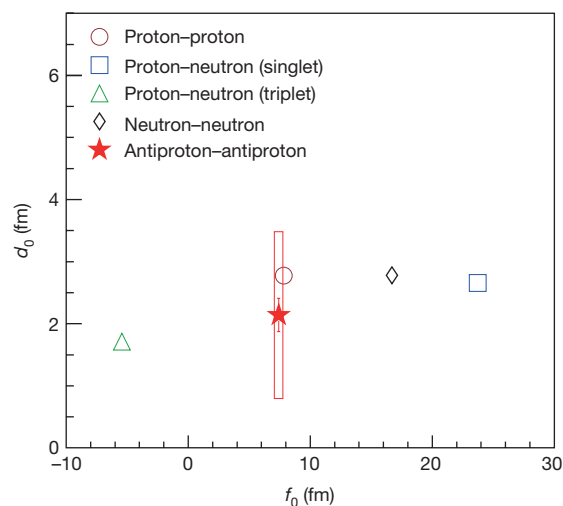


mis-identified  $\Lambda$ s (ref. 22).  $R_{pp}$  and  $R_{p\Lambda}$ , assumed to be the same numerically, are the invariant Gaussian radii<sup>21</sup> from the proton–proton correlation and the proton– $\Lambda$  correlation, respectively.  $x_{pp}$ ,  $x_{p\Lambda}$  and  $x_{\Lambda\Lambda}$ , taken from the THERMINATOR2 model<sup>26</sup>, are the relative contributions from pairs with both daughters from the primary collision, pairs with one daughter from the primary collision and the other one from a  $\Lambda$  decay, and pairs with both daughters from a  $\Lambda$  decay, respectively.

Figure 3 shows the CF for proton–proton pairs (Fig. 3a) and antiproton–antiproton pairs (Fig. 3b), for the 30%–80% centrality class of Au + Au collisions at a centre-of-mass energy of 200 GeV per nucleon pair. The proton–proton CF exhibits a maximum at  $k^* \approx 0.02 \text{ GeV } c^{-1}$  due to the attractive singlet  $s$ -wave interaction between the two detected protons and is consistent with previous measurements<sup>27</sup>. The antiproton–antiproton CF shows a similar structure with the maximum appearing at the same  $k^*$  value. In Fig. 3c, the ratio of the inclusive CF for proton–proton pairs to that of antiproton–antiproton pairs is presented. It is well centred at unity for almost all the  $k^*$  range, except for the region  $k^* < 0.02 \text{ GeV } c^{-1}$ , where the error becomes large. This indicates that the strong interaction is indistinguishable within errors between proton–proton pairs and antiproton–antiproton pairs. By fitting the CF with equation (1), we determine the singlet  $s$ -wave



**Figure 3 | Correlation functions and their ratio.** **a, b,** Correlation functions for proton–proton pairs (**a**) and antiproton–antiproton pairs (**b**). The ratio of the former to the latter is shown in **c**. Errors are statistical only. The fits to the data with equation (1),  $C_{\text{inclusive}}(k^*)$ , are plotted as solid lines, and the term  $1 + x_{pp}[C_{pp}(k^*; R_{pp}) - 1]$  is shown as dashed lines. The  $\chi^2$  per number of degrees of freedom of the fit is 1.66 for **a** and 1.61 for **b**. To take advantage of the existing knowledge on the proton–proton interaction, which is relatively well understood, when fitting the proton–proton correlation,  $f_0$  and  $d_0$  for protons are fixed at values measured from proton–proton elastic-scattering experiments, which are 7.82 fm and 2.78 fm, respectively<sup>29</sup>. When fitting the antiproton–antiproton correlation,  $f_0$  and  $d_0$  are treated as free parameters.



**Figure 4 |  $d_0$  versus  $f_0$  for (anti)nucleon-(anti)nucleon interactions.** The singlet  $s$ -wave scattering length ( $f_0$ ) and the effective range ( $d_0$ ) for the antiproton–antiproton interaction (red star) is plotted together with the  $s$ -wave scattering parameters for other nucleon–nucleon interactions. Here, statistical errors are represented by error bars, while the horizontal uncertainty for  $f_0$  is smaller than the symbol size, and systematic errors are represented by the box. Errors on other measurements<sup>29,30</sup> are of the order of a few per cent, smaller than the symbol size.

scattering length and effective range for the antiproton–antiproton interaction to be  $f_0 = 7.41 \pm 0.19(\text{stat.}) \pm 0.36(\text{sys.}) \text{ fm}$  and  $d_0 = 2.14 \pm 0.27(\text{stat.}) \pm 1.34(\text{sys.}) \text{ fm}$ , respectively. Here stat. and sys. indicate statistical and systematic errors, respectively. The extracted radii for protons ( $R_{pp}$ ) and that for antiprotons ( $R_{\bar{p}\bar{p}}$ ) are  $2.75 \pm 0.01(\text{stat.}) \pm 0.04(\text{sys.}) \text{ fm}$  and  $2.80 \pm 0.02(\text{stat.}) \pm 0.03(\text{sys.}) \text{ fm}$ , respectively.

Figure 4 presents the first measurement of the antiproton–antiproton interaction, together with prior measurements for nucleon–nucleon interactions. Within errors, the  $f_0$  and  $d_0$  for the antiproton–antiproton interaction are consistent with their antiparticle counterparts—the ones for the proton–proton interaction. Our measurements provide parameterization input for describing the interaction among cold-trapped gases of antimatter ions, as in an ultracold environment, where  $s$ -wave scattering dominates and effective-range theory shows that the scattering length and effective range are parameters that suffice to describe elastic collisions. The result provides a quantitative verification of matter–antimatter symmetry in the important and ubiquitous context of the forces responsible for the binding of (anti)nuclei. Possible future improvement of the measurement could be made by reducing the uncertainty from the  $\Lambda$ – $\Lambda$  CF ( $C_{\Lambda\Lambda}(k^*)$ ), which dominates our systematic error, by further accumulation of data. In addition, a similar extraction of  $f_0$  and  $d_0$  could also be repeated with (anti)proton–(anti)proton CF<sup>28</sup> measured at the Large Hadron Collider, where the yield ratio of antiproton to proton is close to unity.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 25 July; accepted 11 September 2015.**

**Published online 4 November; corrected online 18 November 2015**

**(see full-text HTML version for details).**

1. STAR Collaboration. Observation of the antimatter helium-4 nucleus. *Nature* **473**, 353–356 (2011); erratum **475**, 412 (2011).
2. STAR Collaboration. STAR detector overview. *Nucl. Instrum. Methods Phys. Res. A* **499**, 624–632 (2003).
3. Harrison, M., Ludlam, T. & Ozaki, S. RHIC project overview. *Nucl. Instrum. Methods Phys. Res. A* **499**, 235–244 (2003).
4. Hanbury Brown, R. & Twiss, R. Q. A new type of interferometer for use in radio astronomy. *Phil. Mag.* **45**, 663–682 (1954).
5. Yamazaki, T., Ishikawa, K., Kuramashi, Y. & Ukawa, A. Helium nuclei, deuteron, and dineutron in  $2 + 1$  flavor lattice QCD. *Phys. Rev. D* **86**, 074514 (2012).
6. Riotto, A. & Trodden, M. Recent progress in baryogenesis. *Annu. Rev. Nucl. Part. Sci.* **49**, 35–75 (1999).

7. Particle Data Group. Olive, K. A. *et al.* Review of particle physics. *Chin. Phys. C* **38**, 090001, 96–106 (2014).
8. Schellekens, M. *et al.* Hanbury Brown Twiss effect for ultracold quantum gases. *Science* **310**, 648–651 (2005).
9. Kiesel, H., Renz, A. & Hasselbach, F. Observation of Hanbury Brown–Twiss anticorrelations for free electrons. *Nature* **418**, 392–394 (2002).
10. Rom, T. *et al.* Free fermion antibunching in a degenerate atomic Fermi gas released from an optical lattice. *Nature* **444**, 733–736 (2006).
11. Goldhaber, G., Goldhaber, S., Lee, W. & Pais, A. Influence of Bose-Einstein statistics on the antiproton-proton annihilation process. *Phys. Rev.* **120**, 300–312 (1960).
12. Kopylov, G. I. & Podgoretskii, M. I. Interference of two-particle states in elementary-particle physics and astronomy. *Sov. Phys. JETP* **42**, 211–214 (1975).
13. Podgoretskii, M. I. Interference correlations of identical pions. Theory. *Sov. J. Part. Nucl.* **20**, 266–282 (1989).
14. Gyulassy, M., Kauffmann, S. K. & Wilson, L. W. Pion interferometry of nuclear collisions. 1. Theory. *Phys. Rev. C* **20**, 2267–2292 (1979).
15. Boal, H. D., Gelbke, C.-K. & Jennings, B. K. Intensity interferometry in subatomic physics. *Rev. Mod. Phys.* **62**, 553–602 (1990).
16. Lednický, R. Correlation femtoscopy of multiparticle processes. *Phys. Atom. Nucl.* **67**, 72–82 (2004).
17. Lisa, M., Pratt, S., Soltz, R. & Wiedemann, U. Femtoscopy in relativistic heavy ion collisions: two decades of progress. *Ann. Rev. Nucl. Part. Sci.* **55**, 357–402 (2005).
18. Koonin, S. E. Proton pictures of high-energy nuclear collisions. *Phys. Lett. B* **70**, 43–47 (1977).
19. Lednický, R. & Lyuboshitz, V. L. Influence of final-state interaction on correlations of two particles with nearly equal momenta. *Sov. J. Nucl. Phys.* **35**, 770–788 (1982).
20. Lednický, R. Notes on correlation femtoscopy. *Phys. Atom. Nucl.* **71**, 1572–1578 (2008).
21. STAR Collaboration. Proton- $\Lambda$  correlations in central Au+Au collisions at  $\sqrt{s_{NN}} = 200$  GeV. *Phys. Rev. C* **74**, 064906 (2006).
22. STAR Collaboration.  $\Lambda$ - $\Lambda$  correlation function in Au+Au collisions at  $\sqrt{s_{NN}} = 200$  GeV. *Phys. Rev. Lett.* **114**, 022301 (2015).
23. Anderson, M. *et al.* The STAR time projection chamber: a unique tool for studying high multiplicity events at RHIC. *Nucl. Instrum. Methods Phys. Res. A* **499**, 659–678 (2003).
24. STAR Collaboration. Multipair RPCs in the STAR experiment at RHIC. *Nucl. Instrum. Methods Phys. Res. A* **661**, S110–S113 (2012).
25. Kiesel, A., Zbroszczyk, H. & Szymański, M. Extracting baryon-antibaryon strong-interaction potentials from  $p\bar{\Lambda}$  femtoscopic correlation functions. *Phys. Rev. C* **89**, 054916 (2014).
26. Chojnacki, M., Kiesel, A., Florkowski, W. & Broniowski, W. THERMINATOR 2: THERMal heavy ion generator. *Comput. Phys. Commun.* **183**, 746–773 (2012).
27. Pochodzalla, J. *et al.* Two-particle correlations at small relative momenta for  $^{40}\text{Ar}$ -induced reactions on  $^{197}\text{Au}$  at  $E/A = 60$  MeV. *Phys. Rev. C* **35**, 1695–1719 (1987).
28. ALICE Collaboration. One-dimensional pion, kaon, and proton femtoscopy in Pb-Pb collisions at  $\sqrt{s_{NN}} = 2.76$  TeV. Preprint at <http://arxiv.org/abs/1506.07884> (2015).
29. Mathelitsch, L. & VerWest, B. J. Effective range parameters in nucleon-nucleon scattering. *Phys. Rev. C* **29**, 739–746 (1984).
30. Šlaus, I., Akaishi, Y. & Tanaka, H. Neutron-neutron effective range parameters. *Phys. Rep.* **173**, 257–300 (1989).
- R. Esha<sup>32</sup>, O. Evdokimov<sup>33</sup>, O. Eysen<sup>8</sup>, R. Fatemi<sup>2</sup>, S. Fazio<sup>8</sup>, P. Federic<sup>14</sup>, J. Fedorin<sup>3</sup>, Z. Feng<sup>34</sup>, P. Filip<sup>3</sup>, Y. Fisyak<sup>8</sup>, C. E. Flores<sup>18</sup>, L. Fulek<sup>1</sup>, C. A. Gagliardi<sup>20</sup>, D. Garand<sup>35</sup>, F. Geurts<sup>15</sup>, A. Gibson<sup>31</sup>, M. Girard<sup>36</sup>, L. Greiner<sup>25</sup>, D. Grosnick<sup>31</sup>, D. S. Gunaratne<sup>37</sup>, Y. Guo<sup>38</sup>, A. Gupta<sup>11</sup>, S. Gupta<sup>11</sup>, W. Guryan<sup>8</sup>, A. Hamad<sup>7</sup>, H. Hamad<sup>20</sup>, R. Haque<sup>9</sup>, J. W. Harris<sup>17</sup>, L. He<sup>35</sup>, S. Heppelmann<sup>30</sup>, S. Heppelmann<sup>8</sup>, A. Hirsch<sup>35</sup>, G. W. Hoffmann<sup>12</sup>, D. J. Hofman<sup>33</sup>, S. Horvat<sup>17</sup>, B. Huang<sup>32</sup>, H. Z. Huang<sup>32</sup>, X. Huang<sup>32</sup>, P. Huck<sup>34</sup>, T. J. Humanic<sup>19</sup>, G. Igo<sup>32</sup>, W. W. Jacobs<sup>39</sup>, H. Jiang<sup>40</sup>, K. Jiang<sup>38</sup>, E. G. Judd<sup>26</sup>, S. Kabana<sup>7</sup>, D. Kalinkin<sup>6</sup>, K. Kang<sup>23</sup>, K. Kauder<sup>41</sup>, H. W. Ke<sup>8</sup>, D. Keane<sup>7</sup>, A. Kechechyan<sup>3</sup>, Z. H. Khan<sup>33</sup>, D. P. Kikola<sup>36</sup>, I. Kisel<sup>42</sup>, A. Kisel<sup>36</sup>, S. Klein<sup>25</sup>, L. Kochenda<sup>16</sup>, D. D. Koetke<sup>31</sup>, T. Kollegger<sup>42</sup>, L. K. Kosarzewski<sup>36</sup>, A. F. Kraishan<sup>37</sup>, P. Kravtsov<sup>16</sup>, K. Krueger<sup>43</sup>, I. Kulakov<sup>42</sup>, L. Kumar<sup>4</sup>, R. A. Kycia<sup>44</sup>, M. A. C. Lamont<sup>8</sup>, J. M. Landgraf<sup>8</sup>, K. D. Landry<sup>32</sup>, J. Lauret<sup>8</sup>, A. Lebedev<sup>8</sup>, R. Lednický<sup>3</sup>, J. H. Lee<sup>8</sup>, X. Li<sup>37</sup>, Z. M. Li<sup>34</sup>, Y. Li<sup>23</sup>, W. Li<sup>21</sup>, X. Li<sup>8</sup>, C. Li<sup>38</sup>, M. A. Lisa<sup>19</sup>, F. Liu<sup>34</sup>, T. Ljubicic<sup>8</sup>, W. J. Llope<sup>41</sup>, M. Lomnitz<sup>7</sup>, R. S. Longacre<sup>8</sup>, X. Luo<sup>34</sup>, G. L. Ma<sup>21</sup>, R. Ma<sup>8</sup>, Y. G. Ma<sup>21</sup>, L. Ma<sup>21</sup>, N. Magdy<sup>45</sup>, R. Majka<sup>17</sup>, A. Manion<sup>25</sup>, S. Margetis<sup>7</sup>, C. Markert<sup>12</sup>, H. Masui<sup>25</sup>, H. S. Matis<sup>25</sup>, D. McDonald<sup>10</sup>, K. Meenan<sup>18</sup>, N. G. Minaev<sup>29</sup>, S. Mioduszewski<sup>20</sup>, D. Mishra<sup>9</sup>, B. Mohanty<sup>9</sup>, M. M. Mondal<sup>20</sup>, D. A. Morozov<sup>29</sup>, M. K. Mustafa<sup>25</sup>, B. K. Nandi<sup>46</sup>, Md. Nasim<sup>32</sup>, T. K. Nayak<sup>5</sup>, G. Nigmatkulov<sup>16</sup>, L. V. Nogach<sup>29</sup>, S. Y. Noh<sup>40</sup>, J. Novak<sup>47</sup>, S. B. Nurushv<sup>29</sup>, G. Odyniec<sup>25</sup>, A. Ogawa<sup>8</sup>, K. Oh<sup>48</sup>, V. Okorokov<sup>16</sup>, D. Olvitt Jr<sup>37</sup>, B. S. Page<sup>8</sup>, R. Pak<sup>8</sup>, Y. X. Pan<sup>32</sup>, Y. Pandit<sup>33</sup>, Y. Panebratsev<sup>3</sup>, B. Pawlik<sup>44</sup>, H. Pei<sup>34</sup>, C. Perkins<sup>26</sup>, A. Peterson<sup>19</sup>, P. Pile<sup>8</sup>, M. Planinic<sup>49</sup>, J. Pluta<sup>38</sup>, M. A. Lisa<sup>19</sup>, F. Liu<sup>34</sup>, T. Ljubicic<sup>8</sup>, J. Porter<sup>25</sup>, M. Posik<sup>37</sup>, A. M. Poskanzer<sup>25</sup>, J. Putschke<sup>41</sup>, H. Qiu<sup>25</sup>, A. Quintero<sup>7</sup>, S. Ramachandran<sup>2</sup>, R. Raniwala<sup>50</sup>, S. Raniwala<sup>50</sup>, R. L. Ray<sup>12</sup>, H. G. Ritter<sup>25</sup>, J. B. Roberts<sup>15</sup>, O. V. Rogachevskiy<sup>3</sup>, J. L. Romero<sup>18</sup>, A. Roy<sup>5</sup>, L. Ruan<sup>8</sup>, J. Rusnak<sup>14</sup>, O. Rusnakova<sup>13</sup>, N. R. Sahoo<sup>20</sup>, P. K. Sahu<sup>27</sup>, I. Sakrejda<sup>25</sup>, S. Salur<sup>25</sup>, J. Sandweiss<sup>17</sup>, A. Sarker<sup>46</sup>, J. Schambach<sup>12</sup>, R. P. Scharenberg<sup>35</sup>, A. M. Schmah<sup>25</sup>, W. B. Schmidke<sup>8</sup>, N. Schmitz<sup>51</sup>, J. Seger<sup>24</sup>, P. Seyboth<sup>51</sup>, N. Shah<sup>21</sup>, E. Shalaliev<sup>3</sup>, P. V. Shanmuganathan<sup>7</sup>, M. Shao<sup>38</sup>, M. K. Sharma<sup>11</sup>, B. Sharma<sup>4</sup>, W. Q. Shen<sup>21</sup>, S. S. Shi<sup>34</sup>, Q. Y. Shou<sup>21</sup>, E. P. Sichtermann<sup>25</sup>, R. Sikora<sup>1</sup>, M. Simko<sup>14</sup>, M. J. Skoby<sup>39</sup>, N. Smirnov<sup>17</sup>, D. Smirnov<sup>8</sup>, L. Song<sup>10</sup>, P. Sorensen<sup>8</sup>, H. M. Spinka<sup>43</sup>, B. Srivastava<sup>35</sup>, T. D. S. Stanislaus<sup>31</sup>, M. Stepanov<sup>35</sup>, R. Stock<sup>42</sup>, M. Strikhanov<sup>16</sup>, B. Stringfellow<sup>35</sup>, M. Sumner<sup>14</sup>, B. Summa<sup>30</sup>, Z. Sun<sup>22</sup>, X. M. Sun<sup>34</sup>, Y. Sun<sup>38</sup>, X. Sun<sup>25</sup>, B. Surrow<sup>37</sup>, N. Svirida<sup>6</sup>, M. A. Szelezniak<sup>25</sup>, Z. Tang<sup>38</sup>, A. H. Tang<sup>7</sup>, T. Tarnowski<sup>47</sup>, A. Tawfik<sup>45</sup>, J. H. Thomas<sup>25</sup>, A. R. Timmins<sup>10</sup>, D. Tlustý<sup>14</sup>, M. Tokarev<sup>3</sup>, S. Trentalange<sup>32</sup>, R. E. Trible<sup>20</sup>, P. Tribedy<sup>5</sup>, S. K. Tripathy<sup>27</sup>, B. A. Grzegorz<sup>13</sup>, O. D. Tsai<sup>32</sup>, T. Ullrich<sup>8</sup>, D. G. Underwood<sup>43</sup>, I. Upsal<sup>19</sup>, G. Van Buren<sup>8</sup>, G. van Nieuwenhuizen<sup>38</sup>, M. Vandenbroucke<sup>37</sup>, R. Varma<sup>46</sup>, A. N. Vasiliev<sup>29</sup>, R. Vertesi<sup>14</sup>, F. Videbæk<sup>8</sup>, Y. P. Viyogi<sup>5</sup>, S. Vokal<sup>3</sup>, S. A. Voloshin<sup>41</sup>, A. Vossen<sup>39</sup>, G. Wang<sup>32</sup>, H. Wang<sup>8</sup>, J. S. Wang<sup>22</sup>, Y. Wang<sup>34</sup>, Y. Wang<sup>23</sup>, F. Wang<sup>33</sup>, J. C. Webb<sup>8</sup>, G. Webb<sup>8</sup>, L. Wen<sup>32</sup>, G. D. Westfall<sup>47</sup>, H. Wieman<sup>25</sup>, S. W. Wissink<sup>39</sup>, R. Witt<sup>52</sup>, Y. F. Wu<sup>34</sup>, Z. G. Xiao<sup>23</sup>, W. Xie<sup>35</sup>, K. Xin<sup>15</sup>, Y. F. Xu<sup>21</sup>, Q. H. Xu<sup>28</sup>, H. Xu<sup>22</sup>, N. Xu<sup>25</sup>, Z. Xu<sup>8</sup>, Y. Yang<sup>22</sup>, C. Yang<sup>38</sup>, S. Yang<sup>38</sup>, Y. Yang<sup>34</sup>, Q. Yang<sup>38</sup>, Z. Ye<sup>33</sup>, P. Yepes<sup>15</sup>, L. Yi<sup>17</sup>, K. Yip<sup>8</sup>, I.-K. Yoo<sup>48</sup>, N. Yu<sup>34</sup>, H. Zbroszczyk<sup>36</sup>, W. Zha<sup>38</sup>, J. B. Zhang<sup>34</sup>, Z. Zhang<sup>21</sup>, J. Zhang<sup>28</sup>, S. Zhang<sup>21</sup>, X. P. Zhang<sup>33</sup>, J. Zhang<sup>22</sup>, Y. Zhang<sup>38</sup>, J. Zhao<sup>34</sup>, C. Zhong<sup>21</sup>, L. Zhou<sup>38</sup>, X. Zhu<sup>23</sup>, Y. Zoukarnaveva<sup>3</sup> & M. Zyzak<sup>42</sup>

**Acknowledgements** We thank the RHIC Operations Group and RCF at BNL, the NERSC Center at LBNL, the KISTI Center in Korea, and the Open Science Grid consortium for providing resources and support. This work was supported in part by the Office of Nuclear Physics within the US DOE Office of Science, the US NSF, the Ministry of Education and Science of the Russian Federation, NSFC, the MoST of China (973 Programme No. 2014CB845400), CAS, MoST and MoE of China, the Korean Research Foundation, GA and MSM of the Czech Republic, FIAS of Germany, DAE, DST and UGC of India, the National Science Centre of Poland, National Research Foundation, the Ministry of Science, Education and Sports of the Republic of Croatia, and RosAtom of Russia.

**Author Contributions** All authors contributed equally.

**Additional Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to The STAR Collaboration ([star-antiprotonf0d0-l@lists.bnl.gov](mailto:star-antiprotonf0d0-l@lists.bnl.gov)).

## The STAR Collaboration

L. Adamczyk<sup>1</sup>, J. K. Adkins<sup>2</sup>, G. Agakishiev<sup>3</sup>, M. M. Aggarwal<sup>4</sup>, Z. Ahammed<sup>5</sup>, I. Alekseev<sup>6</sup>, J. Alford<sup>7</sup>, A. Aparin<sup>3</sup>, D. Arkhipkin<sup>8</sup>, E. C. Aschenauer<sup>8</sup>, G. S. Averichev<sup>3</sup>, V. Bairathi<sup>9</sup>, A. Banerjee<sup>5</sup>, R. Bellwied<sup>10</sup>, A. Bhasin<sup>11</sup>, A. K. Bhati<sup>4</sup>, P. Bhattarai<sup>12</sup>, J. Bielcik<sup>13</sup>, J. Bielcikova<sup>14</sup>, L. C. Bland<sup>8</sup>, I. G. Bordyuzhin<sup>6</sup>, J. Bouchet<sup>7</sup>, J. D. Brandenburg<sup>15</sup>, A. V. Brandin<sup>16</sup>, I. Bunzarov<sup>3</sup>, J. Butterworth<sup>15</sup>, H. Caines<sup>17</sup>, M. Calderón de la Barca Sánchez<sup>18</sup>, J. M. Campbell<sup>19</sup>, D. Cebra<sup>18</sup>, M. C. Cervantes<sup>20</sup>, I. Chakaberia<sup>8</sup>, P. Chaloupka<sup>13</sup>, Z. Chang<sup>20</sup>, S. Chattopadhyay<sup>5</sup>, J. H. Chen<sup>21</sup>, X. Chen<sup>22</sup>, J. Cheng<sup>23</sup>, M. Cherney<sup>24</sup>, W. Christie<sup>8</sup>, G. Contin<sup>25</sup>, H. J. Crawford<sup>26</sup>, S. Das<sup>27</sup>, L. C. De Silva<sup>24</sup>, R. R. Debbe<sup>8</sup>, T. G. Dedovich<sup>3</sup>, J. Deng<sup>28</sup>, A. A. Derevschikov<sup>29</sup>, B. di Ruzza<sup>8</sup>, L. Didenko<sup>8</sup>, C. Dilks<sup>30</sup>, X. Dong<sup>25</sup>, J. L. Drachenberg<sup>31</sup>, J. E. Draper<sup>18</sup>, C. M. Du<sup>22</sup>, L. E. Dunkelberger<sup>32</sup>, J. C. Dunlop<sup>8</sup>, L. G. Efimov<sup>3</sup>, J. Engelage<sup>26</sup>, G. Eppley<sup>15</sup>,

<sup>1</sup>AGH University of Science and Technology, Cracow 30-059, Poland. <sup>2</sup>University of Kentucky, Lexington, Kentucky, 40506-0055, USA. <sup>3</sup>Joint Institute for Nuclear Research, Dubna, 141 980, Russia. <sup>4</sup>Panjab University, Chandigarh 160014, India. <sup>5</sup>Variable Energy Cyclotron Centre, Kolkata 700064, India. <sup>6</sup>Alkhanov Institute for Theoretical and Experimental Physics, Moscow 117218, Russia. <sup>7</sup>Kent State University, Kent, Ohio 44242, USA. <sup>8</sup>Brookhaven National Laboratory, Upton, New York 11973, USA. <sup>9</sup>National Institute of Science Education and Research, Bhubaneswar 751005, India. <sup>10</sup>University of Houston, Houston, Texas 77204, USA. <sup>11</sup>University of Jammu, Jammu 180001, India. <sup>12</sup>University of Texas, Austin, Texas 78712, USA. <sup>13</sup>Czech Technical University in Prague, FNSPE, Prague, 115 19, Czech Republic. <sup>14</sup>Nuclear Physics Institute AS CR, 250 68 ež/Prague, Czech Republic. <sup>15</sup>Rice University, Houston, Texas 77251, USA. <sup>16</sup>Moscow Engineering Physics Institute, Moscow 115409, Russia. <sup>17</sup>Yale University, New Haven, Connecticut 06520, USA. <sup>18</sup>University of California, Davis, California 95616, USA. <sup>19</sup>Ohio State University, Columbus, Ohio 43210, USA. <sup>20</sup>Texas A&M University, College Station, Texas 77843, USA. <sup>21</sup>Shanghai Institute of Applied Physics, Shanghai 201800, China. <sup>22</sup>Institute of Modern Physics, Lanzhou 730000, China. <sup>23</sup>Tsinghua University, Beijing 100084, China. <sup>24</sup>Creighton University, Omaha, Nebraska 68178, USA. <sup>25</sup>Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. <sup>26</sup>University of California, Berkeley, California 94720, USA. <sup>27</sup>Institute of Physics, Bhubaneswar 751005, India. <sup>28</sup>Shandong University, Jinan, Shandong 250100, China. <sup>29</sup>Institute of High Energy Physics, Protvino 142281, Russia. <sup>30</sup>Pennsylvania State University, University Park, Pennsylvania 16802, USA. <sup>31</sup>Valparaiso University, Valparaiso, Indiana 46383, USA. <sup>32</sup>University of California, Los Angeles, California 90095, USA. <sup>33</sup>University of Illinois at Chicago, Chicago, Illinois 60607, USA. <sup>34</sup>Central China Normal University (HZNU), Wuhan 430079, China. <sup>35</sup>Purdue University, West Lafayette, Indiana 47907, USA. <sup>36</sup>Warsaw University of Technology, Warsaw 00-661, Poland. <sup>37</sup>Temple University, Philadelphia, Pennsylvania 19122, USA. <sup>38</sup>University of Science and Technology of China, Hefei 230026, China. <sup>39</sup>Indiana University, Bloomington, Indiana 47408, USA. <sup>40</sup>Korea Institute of Science and Technology Information, Daejeon 305-701, South Korea. <sup>41</sup>Wayne State University, Detroit, Michigan 48201, USA. <sup>42</sup>Frankfurt Institute for Advanced Studies FIAS, Frankfurt 60438, Germany. <sup>43</sup>Argonne National Laboratory, Argonne, Illinois 60439, USA. <sup>44</sup>Institute of Nuclear Physics PAN, Cracow 31-342, Poland. <sup>45</sup>World Laboratory for Cosmology and Particle Physics (WLCAPP), Cairo 11571, Egypt. <sup>46</sup>Indian Institute of Technology, Mumbai 400076, India. <sup>47</sup>Michigan State University, East Lansing, Michigan 48824, USA. <sup>48</sup>Pusan National University, Pusan 609735, South Korea. <sup>49</sup>University of Zagreb, Zagreb, HR-10002, Croatia. <sup>50</sup>University of Rajasthan, Jaipur 302004, India. <sup>51</sup>Max-Planck-Institut für Physik, Munich 80805, Germany. <sup>52</sup>United States Naval Academy, Annapolis, Maryland 21402, USA.

## METHODS

**Event mixing for non-correlated pairs and the correction for purity.** Non-correlated pairs each consist of two daughter particles. These daughters belong to two events which are carefully chosen so that they have similar event multiplicity and topology. The ratio  $A(k^*)/B(k^*)$  (see above), after being normalized at a large  $k^*$  (at least  $0.25 \text{ GeV } c^{-1}$ ), gives the measured CF,  $C(k^*)_{\text{meas}}$ . Because in practice one cannot select 100% pure (anti)protons, a correction to pairs is applied to obtain the PID-purity-corrected CF:  $C_{\text{PurityCorrected}}(k^*) = \frac{C_{\text{meas}}(k^*) - 1}{\text{PairPurity}(k^*)} + 1$ .

For simplicity, in equation (1) the subscript “meas” is dropped, and elsewhere in this paper, the subscript “PurityCorrected” is dropped.

**The transformation from  $k_{p\Lambda}^*$  and  $k_{\Lambda\Lambda}^*$  to  $k_{pp}^*$ .** The residual CF  $\tilde{C}_{p\Lambda}(k^*)$  in equation (1) is naturally expressed as an integral transformation of the parent CF  $C_{p\Lambda}(k_{p\Lambda}^*)$ . Here  $k_{p\Lambda}^*$  (and  $k^* = k_{pp}^*$ ) is the magnitude of the three-momentum of either particle in the pair rest frame, while in this case for  $k_{pp}^*$ , one of the protons is the decay daughter of  $\Lambda$ . This transformation is done by  $\tilde{C}_{p\Lambda}(k_{pp}^*) = \int C_{p\Lambda}(k_{p\Lambda}^*) T(k_{p\Lambda}^*, k_{pp}^*) dk_{p\Lambda}^*$ , where  $T(k_{p\Lambda}^*, k_{pp}^*)$  is a matrix that transforms  $k_{p\Lambda}^*$  to  $k_{pp}^*$  (ref. 25). The transformation matrix is generated with the THERMINATOR2 model<sup>26</sup> which is a Monte Carlo event generator dedicated to studies of the statistical production of particles in relativistic heavy-ion collisions. The same procedure is also used in the transformation from  $k_{\Lambda\Lambda}^*$  to  $k_{pp}^*$ .

**The calculation of the FSI contribution to the correlation function.** The femtosopic correlations due to the Coulomb FSI between the emitted electron and the residual nucleus in beta decay have been well known for more than 80 years; they reveal themselves in a sensitivity of the Fermi function (an analogue of the CF<sup>31</sup>) to the nuclear radius. Compared with non-interacting particles, the FSI effect in a two-particle system with total spin  $S$  manifests itself in the substitution of the product of plane waves,  $\exp(-ip_1 X_a - ip_2 X_b)$ , by the non-symmetrized Bethe-Salpeter amplitudes  $\Psi_{p_1 p_2}^{S(-)}(X_a, X_b) = \Psi_{p_1 p_2}^{S(+)*}(X_a, X_b)$  (refs 14, 19, 32, 33). For identical particles, the symmetrization requirement in the representation of total pair spin  $S$  takes on the same form for both bosons and fermions: the non-symmetrized amplitude should be substituted by  $[\Psi_{p_1 p_2}^{S(-)}(X_a, X_b) + (-1)^S \Psi_{p_2 p_1}^{S(-)}(X_a, X_b)] / \sqrt{2}$ . In the pair rest frame,  $X_a - X_b = \{t^*, \mathbf{r}^*\}$  and  $p_1 - p_2 = \{\omega_1^* - \omega_2^*, 2\mathbf{k}^*\}$  where  $\omega_i^* = (m_i^2 + k^{*2})^{1/2}$  is the energy of a particle of mass  $m_i$ , and  $t^*$  and  $\mathbf{r}^*$  are the relative emission time and relative separation in the pair rest frame, respectively. In this frame, the non-symmetrized Bethe-Salpeter amplitude at equal emission times ( $t^* = 0$ ) reduces, up to an inessential phase factor, to a stationary solution of the scattering problem,  $\psi_{-k^*}^{S(+)}(\mathbf{r}^*)$ . At small relative momenta,  $k^* \ll 1/r^*$ , this solution can be used in practical calculations with the condition  $|t^*| \ll m r^{*2}$  (refs 19, 32). The equal-time approximation is almost exact in beta decay, and it is usually quite accurate for particles produced in high-energy collisions (to a few per cent in the FSI contribution to CFs of particles even as light as pions<sup>32</sup>). In collisions involving heavy nuclei, the characteristic separation of the emission points,  $\mathbf{r}^*$ , can be considered substantially larger than the range of the strong-interaction potential. The FSI contribution is then independent of the actual potential form and can be calculated analytically with the help of corresponding scattering amplitudes only<sup>34</sup>. At small  $k^*$ , it is basically determined by the  $s$ -wave scattering amplitudes  $f^S(k^*)$  scaled by the separation  $\mathbf{r}^*$  (ref. 19).

**The analytical calculation of the (anti)proton-(anti)proton correlation function.** The (anti)proton-(anti)proton correlation function,  $C_{pp}(k^*; R_{pp})$  in equation (1), can be described by the Lednický and Lyuboshitz analytical model<sup>19</sup>. In this model, the correlation function is calculated as the square of the properly symmetrized wavefunction averaged over the total pair spin  $S$  and the distribution of relative distances ( $\mathbf{r}^*$ ) of particle emission points in the pair rest frame, assuming 1/4 of the singlet and 3/4 of triplet states and a simple Gaussian distribution  $dN/d^3\mathbf{r}^* \approx \exp(-\mathbf{r}^{*2}/(4R_{pp}^2))$ . Starting with the FSI weight of nucleons emitted with the separation  $\mathbf{r}^*$  and detected with the relative momentum  $\mathbf{k}^*$ ,

$$w(\mathbf{k}^*, \mathbf{r}^*) = |\psi_{-k^*}^{S(+)}(\mathbf{r}^*) + (-1)^S \psi_{k^*}^{S(+)}(\mathbf{r}^*)|^2 / 2$$

where  $\psi_{-k^*}^{S(+)}(\mathbf{r}^*)$  is the equal-time ( $t^* = 0$ ) reduced Bethe-Salpeter amplitude which can be approximated by the outer solution of the scattering problem<sup>19,35</sup>. This is

$$\psi_{-k^*}^{S(+)}(\mathbf{r}^*) = e^{i\delta_c} \sqrt{A_c(\eta)} \left[ e^{-ik^* r^*} F(-i\eta, 1, i\xi) + f_c(k^*) \frac{\tilde{G}(\rho, \eta)}{r^*} \right]$$

where  $\eta = (k^* a_c)^{-1}$ ,  $a_c = 57.5 \text{ fm}$  is the Bohr radius for two protons,  $\rho = k^* r^*$ ,  $\xi = k^* \mathbf{r}^* \cdot \rho$ ,  $A_c(\eta)$  is the Coulomb penetration factor given by  $A_c(\eta) = 2\pi\eta [\exp(2\pi\eta) - 1]^{-1}$ ,  $F$  is the confluent hypergeometric function,  $\tilde{G}(\rho, \eta) = \sqrt{A_c(\eta)} [G_0(\rho, \eta) + iF_0(\rho, \eta)]$  is a combination of the regular ( $F_0$ ) and singular ( $G_0$ )  $s$ -wave Coulomb functions,

$$f_c(k^*) = \left[ \frac{1}{f_0} + \frac{1}{2} d_0 k^{*2} - \frac{2}{a_c} h(\eta) - ik^* A_c(\eta) \right]^{-1}$$

is the  $s$ -wave scattering amplitude renormalized by the Coulomb interaction, and  $h(\eta) = \eta^2 \sum_{n=1}^{\infty} [n(n^2 + \eta^2)]^{-1} - C - \ln |\eta|$  (here  $C \doteq 0.5772$  is the Euler constant). The dependence of the scattering parameters on the total pair spin  $S$  is omitted since only the singlet ( $S=0$ )  $s$ -wave FSI contributes in the case of identical nucleons. The theoretical CF at a given  $k^*$  can be calculated as the average FSI weight  $\langle w(\mathbf{k}^*, \mathbf{r}^*) \rangle$  obtained from the separation  $\mathbf{r}^*$ , simulated according to the Gaussian law, and the angle between the vectors  $\mathbf{k}^*$  and  $\mathbf{r}^*$ , simulated according to a uniform cosine distribution. This CF is subject to the integral correction<sup>19</sup>  $-A_c(\eta) |f_c(k^*)|^2 d_0 / (8\sqrt{\pi} R_{pp}^3)$  due to the deviation of the outer solution from the true wavefunction in the inner potential region. In addition, in Au + Au collisions the emitting source has a net positive charge, and it influences the CF differently for proton and antiproton pairs. This effect is included in the consideration according to refs 32, 33.

**Systematic uncertainties.** The systematic uncertainties include variations due to track-wise and pair-wise cuts, the uncertainty in describing the  $C_{p\Lambda}$  correlation function<sup>36</sup>, and the uncertainty from the  $C_{\Lambda\Lambda}$  measurement. The latter dominates the systematic error of  $d_0$  and  $f_0$ , and it affects  $d_0$  more than it does  $f_0$  because the shape of the CF is sensitive to  $d_0$ , in particular at low  $k^*$ . As a consistency check, when fitting the proton-proton CF, both  $f_0$  and  $d_0$  are also allowed to vary freely, and the fitted  $f_0$  and  $d_0$  agree with the results from fitting the antiproton-antiproton CF. Assuming the measurements from different systematic checks follow a uniform distribution, the final systematic error is given by (maximum – minimum) /  $\sqrt{12}$ . In our calculations, we consider the two-proton wavefunction, taking into account the Coulomb interaction between point-like protons in all orbital angular momentum waves and the strong interaction in the  $s$ -wave only. We neglect the small non-Coulomb electromagnetic contributions due to magnetic interactions, vacuum polarization, and the finite proton size<sup>29,37,38</sup>. This approximation changes the scattering parameters at the level of a few per cent<sup>29,37,38</sup>. The decomposition of systematics from our analysis can be found in Extended Data Table 1.

**Sample size.** No statistical methods were used to predetermine sample size.

- Lednický, R. Femtosopic correlations in multiparticle production and beta-decay. *Braz. J. Phys.* **37**, 939–946 (2007).
- Lednický, R. Finite-size effect on two-particle production in continuous and discrete spectrum. *Phys. Part. Nucl.* **40**, 307–352 (2009).
- Erazmus, B. et al. Influence of the emitting nucleus on the light-particle correlation function. *Nucl. Phys. A* **583**, 395–400 (1995).
- Gmitro, M., Kvasil, J., Lednický, R. & Lyuboshitz, V. L. On the sensitivity of nucleon-nucleon correlations to the form of short-range potential. *Czech. J. Phys. B* **36**, 1281–1287 (1986).
- Landau, L. D. & Lifshitz, E. M. *Kvantovaya Mekhanika: Nerelevativistskaya Teoriya* 3rd edn 585–685 (Nauka, 1974); Landau, L. D. & Lifshitz, E. M. *Quantum Mechanics: Non-relativistic theory* 3rd edn (Pergamon, 2013) [transl.].
- Bodmer, A. R. & Usmani, Q. N. Coulomb effects and charge symmetry breaking for the  $A=4$  hypernuclei. *Phys. Rev. C* **31**, 1400–1411 (1985).
- Heller, L. Interaction of two nucleons at low energies. *Rev. Mod. Phys.* **39**, 584–590 (1967).
- Bergervoet, J. R., van Campen, P. C., van der Sanden, W. A. & de Swart, J. J. Phase shift analysis of 0–30 MeV pp scattering data. *Phys. Rev. C* **38**, 15–50 (1988).



Extended Data Table 1 | The decomposition of systematic errors

	$\Delta f_0 (\pm \text{fm})$	$\Delta d_0 (\pm \text{fm})$	$\Delta R_{\bar{p}\bar{p}} (\pm \text{fm})$	$\Delta R_{pp} (\pm \text{fm})$
experimental cuts	0.14	0.33	0.01	0.03
uncertainty of p- $\Lambda$ CF	0.17	0.19	0.03	0.01
uncertainty of $\Lambda$ - $\Lambda$ CF	0.36	1.34	0.03	0.03
THERMINATOR2 model	0.07	0.09	< 0.01	< 0.01

The table presents systematic uncertainties for  $f_0$  and  $d_0$  for antiproton–antiproton interaction, and  $R$  for both proton–proton and antiproton–antiproton interaction. Errors are listed separately by their sources. Assuming the measurements ( $f_0$ ,  $d_0$  and  $R$ ) from different systematic checks follow a uniform distribution, the systematic error is given by (maximum measurement – minimum measurement)/ $\sqrt{12}$ .

# Nanostructure surveys of macroscopic specimens by small-angle scattering tensor tomography

Marianne Liebi<sup>1</sup>, Marios Georgiadis<sup>2</sup>, Andreas Menzel<sup>1</sup>, Philipp Schneider<sup>3</sup>, Joachim Kohlbrecher<sup>1</sup>, Oliver Bunk<sup>1</sup> & Manuel Guizar-Sicairos<sup>1</sup>

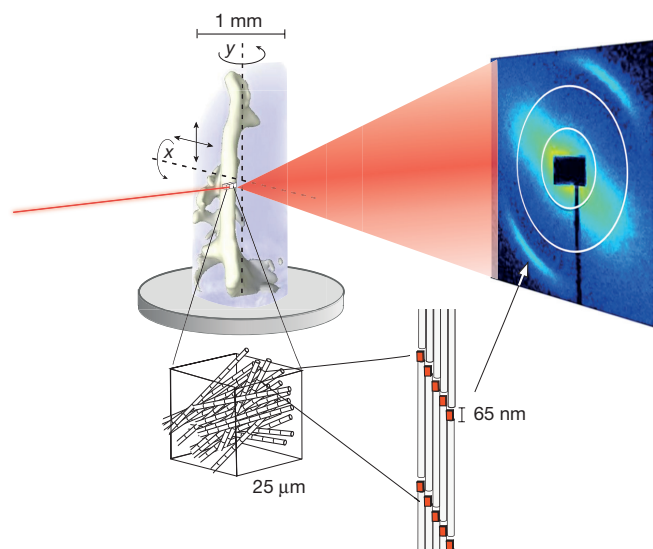
The mechanical properties of many materials are based on the macroscopic arrangement and orientation of their nanostructure. This nanostructure can be ordered over a range of length scales. In biology, the principle of hierarchical ordering is often used to maximize functionality, such as strength and robustness of the material, while minimizing weight and energy cost. Methods for nanoscale imaging provide direct visual access to the ultrastructure (nanoscale structure that is too small to be imaged using light microscopy), but the field of view is limited and does not easily allow a full correlative study of changes in the ultrastructure over a macroscopic sample. Other methods of probing ultrastructure ordering, such as small-angle scattering of X-rays or neutrons, can be applied to macroscopic samples; however, these scattering methods remain constrained to two-dimensional specimens<sup>1–4</sup> or to isotropically oriented ultrastructures<sup>5–7</sup>. These constraints limit the use of these methods for studying nanostructures with more complex orientation patterns, which are abundant in nature and materials science. Here, we introduce an imaging method that combines small-angle scattering with tensor tomography to probe nanoscale structures in three-dimensional macroscopic samples in a non-destructive way. We demonstrate the method by measuring the main orientation and the degree of orientation of nanoscale mineralized collagen fibrils in a human trabecula bone sample with a spatial resolution of 25 micrometres. Symmetries within the sample, such as the cylindrical symmetry commonly observed for mineralized collagen fibrils in bone<sup>8–10</sup>, allow for tractable sampling requirements and numerical efficiency. Small-angle scattering tensor tomography is applicable to both biological and materials science specimens, and may be useful for understanding and characterizing smart or bio-inspired materials. Moreover, because the method is non-destructive, it is appropriate for *in situ* measurements and allows, for example, the role of ultrastructure in the mechanical response of a biological tissue or manufactured material to be studied.

The arrangement of the nanoscale and microscale building blocks of materials over extended regions has implications for the functional properties of many such materials. This is particularly the case for many biological structures, where hierarchically ordered organization ensures the required properties of the structure; for example, the high mechanical stability of bone is necessary for structural support of the body, force transmission for locomotion, and mechanical protection of vital organs<sup>11,12</sup>. High-resolution imaging techniques exist that can directly image nanoscale structures<sup>13,14</sup>. Aside from the individual importance of such information, an overview of how the nanoscale structure changes over macroscopic length scales is pivotal to understanding the function of hierarchically organized tissues and materials. In the case of human bone, the structural elements of interest are the mineralized collagen fibrils, and their orientation is one of the factors that determine bone strength at different structural levels<sup>15–17</sup>.

Experimental techniques exist to probe such structural orientations, by using light, X-rays, and electrons, which use direction-dependent polarization<sup>18–20</sup>, small- or wide-angle scattering<sup>1,18</sup> or diffraction<sup>21,22</sup>, rather than directly imaging the nanostructures. These techniques cover a large field of view of thin sections or of the sample surface. However, studying 3D samples without the need for sectioning can be beneficial; cutting can influence the structures at the surface and, for some samples, cutting is not feasible or applicable at all.

X-ray techniques are adept at probing the inside of macroscopic objects, owing to their high penetration depth, even in dense materials such as mineralized bone tissue. X-ray photons scatter where changes in electron density occur. The scattering angle, or momentum transfer  $q$ , is inversely related to the characteristic sizes of the structure under investigation. Thus, at high scattering angles, atomic length scales can be resolved (using wide-angle X-ray scattering, WAXS, or X-ray diffraction, XRD), whereas at small angles, scattering from structures at the nanometre scale can be detected (using small-angle X-ray scattering, SAXS). Ultrastructure with a high degree of orientation results in an anisotropic scattering pattern, which can be captured with a two-dimensional (2D) detector, and the orientation analysed in each  $q$  range separately<sup>3</sup>. Spatially resolved SAXS, where a specimen is scanned through an X-ray pencil beam, is a valuable tool for mapping ultrastructural features in macroscopic samples<sup>1–3,8</sup>. However, the samples must be thin, or at least of uniform structure in the direction of the incoming beam, to avoid losing spatial information as a result of averaging along the beam path. There have been efforts<sup>5–7,23,24</sup> to combine SAXS or WAXS with computed tomography by rotating the sample with respect to the X-ray beam and scanning at different angular steps. In these cases, the reconstruction was carried out by integrating the scattered intensity over a  $q$  range of interest and applying standard computed-tomography reconstruction algorithms, which use filtered backprojection or algebraic reconstruction techniques. This reconstruction procedure is applicable only if the scattering is invariant with respect to sample rotation, as is the case for isotropic scatterers or tissues and materials with structural symmetry around the rotation axis. For more general, anisotropically oriented ultrastructures, a reconstruction of the full three-dimensional (3D) reciprocal-space map is needed to provide a representation of the distribution of the nanoscale structure. Thus, in each subvolume of the structure (a voxel), a tensor needs to be reconstructed, in contrast to a scalar value as is reconstructed with X-ray absorption tomography. Tensor tomography is applied in magnetic resonance imaging, where the orientation-dependent magnetic relaxation is used to measure the diffusion of water in tissue (diffusion tensor imaging)<sup>25</sup>. Directional dark-field imaging has been successfully combined with tensor tomography to investigate structures and orientations at the micrometre scale<sup>26</sup>. We developed a new algorithm, which combines small-angle scattering with tensor tomography, allowing nanoscale structures to be probed and enabling a reconstruction of the 3D reciprocal-space map for each  $q$  range individually.

<sup>1</sup>Paul Scherrer Institut, 5232 Villigen PSI, Switzerland. <sup>2</sup>Institute for Biomechanics, ETH Zurich, 8093 Zurich, Switzerland. <sup>3</sup>Bioengineering Science Research Group, Faculty of Engineering and the Environment, University of Southampton, Southampton SO17 1BJ, UK.



**Figure 1 | Schematic of the experimental set-up for SAS tensor tomography.** The sample is scanned through an X-ray pencil beam ( $25 \times 25 \mu\text{m}^2$ ) at different sample orientations while a 2D detector is recording the SAS pattern for each scanning point. In addition to rotation around the tomographic axis ( $y$ ), different tilt angles around the  $x$  axis are measured. For the tomographic reconstruction, the sample volume is subdivided into small subvolumes (voxels), the size of which are defined by the width of the X-ray pencil beam. The detector frame probes ultrastructural length scales from a few nanometres up to a few hundred nanometres. This allows us, for instance, to probe mineralized collagen fibrils, as illustrated. The distinct Bragg reflection highlighted by an arrow originates from the regular arrangement of the approximately 65-nm gaps in the collagen fibrils, which contain mineral crystals. The scattering of the fan-shaped profile perpendicular to it originates from the shape, size and lateral arrangement of the mineral crystals, and is affected in the investigated range (indicated by white circles) by the lateral packing of the collagen fibrils<sup>9</sup>.

Figure 1 shows a schematic of the experimental set-up and the principle of the method, which we refer to as small-angle scattering (SAS) tensor tomography. It allows tissues and materials with anisotropically oriented nanostructure to be studied at 3D spatial resolutions limited only by the size of the X-ray beam, which is typically at the micrometre scale.

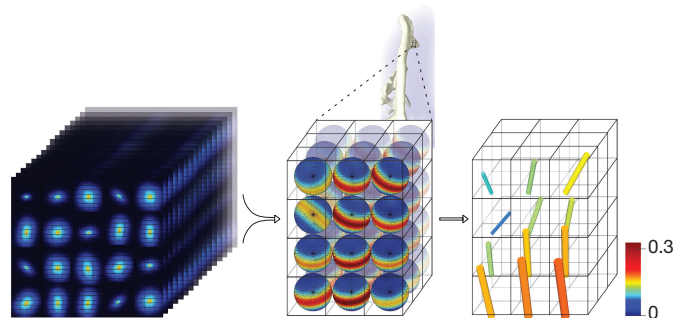
To demonstrate the method, we present a study of a trabecular bone specimen with a size of about  $1 \times 1 \times 2.5 \text{ mm}^3$  extracted from a human vertebra. The scattering angles covered by the detector correspond to structures in the range of 5.4 nm to 300 nm. This range includes the multi-scale structural characteristics of bone tissue, including the arrangement of the hydroxyapatite mineral crystal platelets with sizes of approximately  $3 \times 25 \times 50 \text{ nm}^3$ , their intra-fibrillar spacing (approximately 65 nm), and the diameters of collagen fibrils (50–200 nm) (refs 9, 12, 27). Most of the SAXS signal originates from the contrast in electron density between the mineral crystal platelets and the surrounding lower-electron-density materials such as collagen fibrils and water<sup>1</sup>. It has been shown that the orientation of crystals, collagen fibrils, fibril bundles and fibres are closely related<sup>2,8,12,17</sup>. On the basis of this result, any of these structures can be probed to estimate the orientation of mineralized collagen fibrils in bone.

The sample is scanned through a pencil X-ray beam, which had dimensions of  $25 \times 25 \mu\text{m}^2$  for this experiment. In each scanning point, a 2D scattering pattern is recorded, which is a sum of the scattering of all voxels in the beam path. Similar to first-generation tomography, this scan is repeated for each orientation around the tomographic ( $y$ ) axis. If the step size of the scanning and the angular spacing are chosen appropriately, the spatial resolution of the tomogram is limited only by the size of the X-ray beam. Accordingly, the scanning step size was  $25 \mu\text{m}$  and the angular spacing was  $4.5^\circ$  between  $0^\circ$  and  $180^\circ$ . In

contrast to the X-ray absorption probed in conventional tomography, the 2D scattering pattern that emerges from each probed voxel depends on the relative orientation of the underlying nanostructure with respect to the X-ray beam, and thus changes with sample orientation. Hence, for the reconstruction of the 3D reciprocal-space map in each voxel, additional data need to be collected, which required multiple tilt angles of the tomographic axis. Here, data were collected at six different tilt angles of the tomographic axis, resulting in a total of 240 measured projections, each of which was measured by scanning  $122 \times 55$  positions, resulting in 1.6 million scattering patterns with a total exposure time of 22.5 h. For each scattering pattern, we reduced the data by azimuthal integration over 16 sectors and radial integration over the  $q$  range corresponding to distances from 85 nm to 165 nm, as indicated by the white circles in the detector frame in Fig. 1.

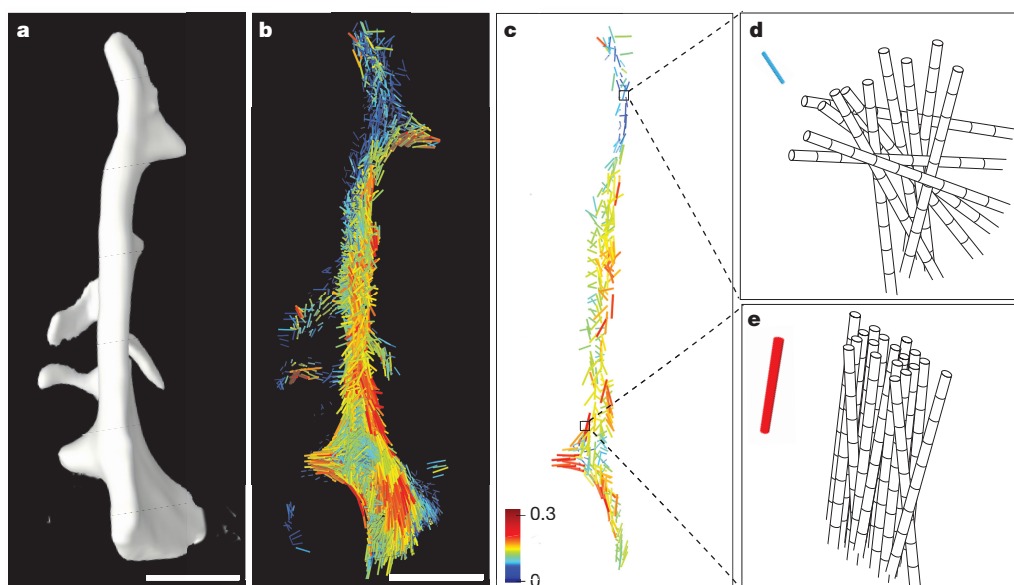
For reconstruction of the orientation, we model the 3D reciprocal-space map with a series of spherical harmonics  $Y_l^m$  of degree  $l$  and order  $m$ , which provide a complete basis for the intensity distribution on the sphere corresponding to a fixed  $q$  range. Spherical harmonics have been used in texture analysis to reconstruct orientation distribution functions on the basis of pole-figure analysis from XRD measurements<sup>28,29</sup>. One advantage of using spherical harmonics is that known symmetries can be enforced by choosing the appropriate degree  $l$  and order  $m$ . For the trabecular bone studied, we assume that the mineral platelet surface normal is randomly oriented in the plane perpendicular to the fibrils, so we expect a point symmetry around  $q = 0$  and a rotational symmetry around one axis in the 3D reciprocal-space map in the  $q$  range investigated. Therefore, we use here only the spherical harmonic functions with even degree  $l$  and zero order  $m$ . An optimization algorithm is used to minimize the error between the modelled intensity of voxels in the beam path projected to the detector plane and the measured intensity for each scanning point under all rotation angles. Figure 2 visualizes how all measured scattering patterns are used to reconstruct the reciprocal-space map from which the orientation of the nanostructure is determined.

The main orientation in the 3D reciprocal-space map is defined by the relative orientation of the zenith of the spherical harmonic functions with respect to the coordinate system of the sample. Thus the main orientation in each voxel is given by polar and azimuthal spherical coordinate angles, which are optimization parameters of the reconstruction. The degree of orientation is calculated as the ratio between the anisotropic scattering (which is described by spherical harmonic functions with degree  $l > 0$ ) and the total scattering. Assuming co-alignment between collagen fibrils and mineral crystals<sup>2,8,12,17</sup>, the main orientation of the scattering in the investigated  $q$  range reveals the



**Figure 2 | Data processing.** More than a million SAXS patterns (left) are used to reconstruct the 3D reciprocal-space map for each voxel, which is modelled using spherical harmonics (middle panel shows  $4 \times 3 \times 3$  voxels) and provides a representation of the nanoscale structure distribution. From the reconstruction, the main ultrastructure orientation as well as the degree of orientation is determined. The former is visually represented by the orientation of the cylinder (right); the latter is illustrated both by the colour (indicating the ratio of anisotropic scattering to total scattering; see colour scale) and the length of the cylinders.





**Figure 3 | Orientation of collagen fibrils within a human trabecular bone sample.** **a**, Computed-tomography reconstruction obtained from the transmitted intensity using standard filtered backprojection. **b**, Orientation of bone ultrastructure as determined using SAS tensor tomography (see also Supplementary Video 1). **c**, One tomographic slice of the reconstruction shown in **b**. The cylinder orientations represent the main orientation of collagen fibrils in the corresponding voxel as

schematically depicted in **d** and **e**. The degree of orientation is represented by both the colour (indicating the ratio of anisotropic scattering to total scattering) and the length of the cylinders, where a low degree of orientation (blue) means a low degree of alignment of the collagen fibrils (**d**), and a high degree of orientation (red) means the collagen fibrils are well aligned with respect to each other (**e**). Scale bars in **a** and **b** correspond to 0.5 mm.

spatial arrangement of the mineralized collagen fibrils. The anisotropic scattering of bone is related to the fraction of aligned mineralized fibrils on average over a certain voxel. Hence, the higher the degree of orientation, the higher the ratio of aligned fibrils compared to randomly oriented fibrils<sup>10</sup> (see Fig. 3). The degree of orientation is calculated directly from the coefficients of the spherical harmonic functions, which are additional optimization parameters of the reconstruction; see Methods for details. In the visualization of the reconstructed ultrastructure the degree of orientation is represented both by the colour and the length of the cylinders (Fig. 2, right).

Simultaneous to the SAXS patterns, the X-ray absorption was measured with a photo diode mounted on the beam stop blocking the direct beam in front of the detector. The corresponding computed-tomographic reconstruction is shown in Fig. 3a. The result of SAS tensor tomography on the human trabecular bone sample is shown in Fig. 3b as a 3D visualization (see also Supplementary Video 1). One tomographic slice of the reconstruction is shown in Fig. 3c. Domains with a low degree of orientation (blue) and a high degree of orientation (red) are identified.

The presence of domains with a size of several tens of micrometres and the observation that high degrees of orientation are found in places with higher curvature in the trabecula is in agreement with previous findings for trabecular bone<sup>4</sup>. As expected, in the curved regions, the collagen fibrils follow closely the trabecular bone microstructure.

The technique presented here can be tuned to specific characteristics of the samples under investigation. For example, the range of sizes of the ultrastructure probed can be varied by changing the analysed scattering angles, that is, the  $q$  range. One advantage of the method is that the angular scattering range of interest is digitally extracted from the scattering frames such that different length scales can be probed simultaneously and reconstructed from the same data set. Independent of the  $q$  range, the real-space resolution of the spatial mapping can be adapted by the size of the X-ray beam and the scanning step size. The analysis algorithm would also allow the reconstruction of the full  $q$  range, similar to what has been shown for SAXS tomography for isotropic scatterers<sup>30</sup>. This enables the size and form of the scattering objects to be determined, as is done using standard SAS analysis.

The method is not limited to the symmetries apparent in the example of trabecular bone. Spherical harmonics provide a complete basis for modelling the 3D reciprocal space, and azimuthal structure can be captured by adding azimuthal orders  $m$  into the optimization. Yet, prior knowledge of symmetry of the scattering object can be used to reduce the complexity of the optimization problem and, potentially, the number of projections needed to reconstruct the structure. Limits on the complexity of a 3D reciprocal-space map that can be recovered with the method presented here and rational sampling requirements remain to be explored. Nevertheless, the flexibility of resolution both in real and reciprocal space and the non-destructiveness of SAS tensor tomography make this imaging technique widely applicable in biological and materials science, providing information complementary to nanoscale imaging, which will support modelling and improve understanding of biological tissues and complex materials.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 3 June; accepted 29 September 2015.**

1. Fratzl, P., Jakob, H. F., Rinnerthaler, S., Roschger, P. & Klaushofer, K. Position-resolved small-angle X-ray scattering of complex biological materials. *J. Appl. Cryst.* **30**, 765–769 (1997).
2. Rinnerthaler, S. *et al.* Scanning small angle X-ray scattering analysis of human bone sections. *Calcif. Tissue Int.* **64**, 422–429 (1999).
3. Bunk, O. *et al.* Multimodal x-ray scatter imaging. *New J. Phys.* **11**, 123086 (2009).
4. Georgiadis, M. *et al.* 3D scanning SAXS: a novel method for the assessment of bone ultrastructure orientation. *Bone* **71**, 42–52 (2015).
5. Jensen, T. H. *et al.* Molecular X-ray computed tomography of myelin in a rat brain. *Neuroimage* **57**, 124–129 (2011).
6. Schroer, C. G. *et al.* Mapping the local nanostructure inside a specimen by tomographic small-angle x-ray scattering. *Appl. Phys. Lett.* **88**, 164102 (2006).
7. Álvarez-Murga, M., Bleuet, P. & Hodeau, J.-L. Diffraction/scattering computed tomography for three-dimensional characterization of multi-phase crystalline and amorphous materials. *J. Appl. Cryst.* **45**, 1109–1124 (2012).
8. Seidel, R. *et al.* Synchrotron 3D SAXS analysis of bone nanostructure. *Bioinspir. Biomim.* **1**, 123–131 (2012).
9. Giannini, C. *et al.* Scanning SAXS-WAXS microscopy on osteoarthritis-affected bone – an age-related study. *J. Appl. Cryst.* **47**, 110–117 (2014).
10. Pabisch, S., Wagermaier, W., Zander, T., Li, C. H. & Fratzl, P. in *Methods in Enzymology* Vol. 532 (ed. De Yoreo, J. J.) 391–413 (Elsevier, 2013).

11. Currey, J. D. *Bones: Structure and Mechanics* (Princeton Univ. Press, 2002).
12. Fratzl, P. & Weinkamer, R. Nature's hierarchical materials. *Prog. Mater. Sci.* **52**, 1263–1334 (2007).
13. Schneider, P., Meier, M., Wepf, R. & Muller, R. Towards quantitative 3D imaging of the osteocyte lacuno-canalicular network. *Bone* **47**, 848–858 (2010).
14. Holler, M. *et al.* X-ray ptychographic computed tomography at 16 nm isotropic 3D resolution. *Sci. Rep.* **4**, 3857 (2014).
15. Martin, R. B. & Ishida, J. The relative effects of collagen fiber orientation, porosity, density, and mineralization on bone strength. *J. Biomech.* **22**, 419–426 (1989).
16. Riggs, C. M., Vaughan, L. C., Evans, G. P., Lanyon, L. E. & Boyde, A. Mechanical implications of collagen fibre orientation in cortical bone of the equine radius. *Anat. Embryol.* **187**, 239–248 (1993).
17. Granke, M. *et al.* Microfibril orientation dominates the microelastic properties of human bone tissue at the lamellar length scale. *PLoS ONE* **8**, e58043 (2013).
18. Giannini, C. *et al.* Correlative light and scanning X-ray scattering microscopy of healthy and pathologic human bone sections. *Sci. Rep.* **2**, 435 (2012).
19. Zhao, Q. & Wagner, H. D. Raman spectroscopy of carbon-nanotube-based composites. *Phil. Trans. R. Soc. London Ser. A* **362**, 2407–2424 (2004).
20. Bi, X., Li, G., Doty, S. B. & Camacho, N. P. A novel method for determination of collagen orientation in cartilage by Fourier transform infrared imaging spectroscopy (FT-IRIS). *Osteoarthritis Cartilage* **13**, 1050–1058 (2005).
21. Rauch, E. F. *et al.* Automated nanocrystal orientation and phase mapping in the transmission electron microscope on the basis of precession electron diffraction. *Z. Krist.* **225**, 103–109 (2010).
22. Heidebach, F., Riekel, C. & Wenk, H.-R. Quantitative texture analysis of small domains with synchrotron radiation X-rays. *J. Appl. Cryst.* **32**, 841–849 (1999).
23. Feldkamp, J. M. *et al.* Recent developments in tomographic small-angle X-ray scattering. *Phys. Status Solidi A* **206**, 1723–1726 (2009).
24. Ludwig, W., Schmidt, S., Lauridsen, E. M. & Poulsen, H. F. X-ray diffraction contrast tomography: a novel technique for three-dimensional grain mapping of polycrystals. I. Direct beam case. *J. Appl. Cryst.* **41**, 302–309 (2008).
25. Bassar, P. J., Mattiello, J. & Lebihan, D. Estimation of the effective self-diffusion tensor from the NMR spin echo. *J. Magn. Reson. B.* **103**, 247–254 (1994).
26. Malecki, A. *et al.* X-ray tensor tomography. *EPL (Europhys. Lett.)* **105**, 38002 (2014).
27. Gourrier, A. *et al.* Scanning small-angle X-ray scattering analysis of the size and organization of the mineral nanoparticles in fluorotic bone using a stack of cards model. *J. Appl. Cryst.* **43**, 1385–1392 (2010).
28. Roe, R. J. Description of crystallite orientation in polycrystalline materials. III. General solution to pole figure inversion. *J. Appl. Phys.* **36**, 2024–2031 (1965).
29. Bunge, H. J. & Roberts, W. T. Orientation distribution, elastic and plastic anisotropy in stabilized steel sheet. *J. Appl. Cryst.* **2**, 116–128 (1969).
30. Jensen, T. H. *et al.* Brain tumor imaging using small-angle x-ray scattering tomography. *Phys. Med. Biol.* **56**, 1717–1726 (2011).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank M. Holler and J. Raabe for their help in sample preparation and A. Diaz, F. Schaff and M. Bech for discussions. M.G. was supported by the ETH Research Grant ETH-39 11-1. The vertebral specimen was provided by W. Schmölz, Department for Trauma Surgery, Innsbruck Medical University, Innsbruck, Austria.

**Author Contributions** M.L., M.G., A.M., P.S., O.B., and M.G.-S. conceived the research project. M.G. prepared the sample. M.L., M.G., and M.G.-S. carried out the X-ray experiments. M.L. developed the data analysis framework with support from O.B. and M.G.-S. Results were interpreted by M.L., M.G., P.S., and J.K. M.L. wrote the manuscript with contributions from all authors.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.L. ([marianne.liebi@psi.ch](mailto:marianne.liebi@psi.ch)) or M.G.-S. ([manuel.guizar-sicairos@psi.ch](mailto:manuel.guizar-sicairos@psi.ch)).

## METHODS

**Sample preparation.** A trabecula was extracted from a T12 human vertebra from a 73-year-old man and cleaned from soft tissue. The trabecula (approximately  $1 \times 1 \times 2.5 \text{ mm}^3$  in size) was isolated using a scalpel and embedded into polymethyl methacrylate (PMMA). The human vertebra had been obtained from the Department of Anatomy, Histology, and Embryology at the Innsbruck Medical University, Innsbruck, Austria, with the written consent of the donor according to Austrian law. All subsequent procedures were in accordance to Swiss law, the Guideline on Bio-Banking of the Swiss Academy of the Medical Sciences (2006) and the Swiss ordinance 814.912 (2012) on the contained use of organisms.

**Experiment.** Experiments were carried out at the cSAXS beamline (X12SA) of the Swiss Light Source, Paul Scherrer Institut, Switzerland. The 12.4-keV X-ray beam was defined by a fixed-exit double-crystal Si(111) monochromator to an energy bandwidth of about  $2 \times 10^{-4}$  and focused to a beam size of approximately  $25 \times 25 \mu\text{m}^2$ . Horizontal (sagittal) focusing was achieved by bending the second monochromator crystal and vertical (meridional) focusing by bending a Rh coated mirror. The sample was mounted on a goniometer on two perpendicular rotation stages on top of a two-axis scanning stage. A 7-m-long flight tube under vacuum was inserted between the sample and the detector to minimize air scattering and X-ray absorption. Scattering patterns were collected with a PILATUS 2M detector<sup>31</sup> placed 7.15-m downstream of the sample. Simultaneously, the transmitted beam was measured with a diode mounted on a beam stop blocking the direct beam before the exit window of the flight tube. 40 projections at tilt angles between  $0^\circ$  and  $180^\circ$  around the tomographic rotation axis were collected for 6 different tilt angles of the rotation axis between  $-30^\circ$  and  $45^\circ$ , with 6,710 scan points each. This resulted in a total of 1,610,400 scattering patterns and a total exposure time of 22.5 h, with an individual frame exposure time of 50 ms. Each voxel was exposed for 12 s in total. The dose  $d$  was estimated using  $d = \mu N_0 \epsilon / p$  (ref. 32), where  $\mu/p$  is the mass-attenuation coefficient ( $37.5 \text{ cm}^2 \text{ g}^{-1}$  for bone at 12.4 keV; ref. 33),  $N_0$  is the number of incident photons per unit area, and  $\epsilon$  is the photon energy. Thus the dose imparted on the bone specimen is about  $2.9 \times 10^7 \text{ Gy}$ . To assure that no structural changes were induced by radiation damage, the projection at zero rotation around both axes was repeated at the beginning, in the middle, and at the end of the experiment. In future experiments with increased resolution or on specimens prone to radiation damage, the dose could be reduced by using fewer tilt angles, lower photon counting statistics and a higher X-ray energy than in this proof-of-concept experiment.

The total time for the experiment was 35.5 h, which includes an overhead of 13 h due to motor movements. This will be substantially reduced by faster sample stages with precise positioning, which will also enable higher spatial resolution. Because the exposure time was limited by detector data transfer and not by photon flux, we expect that the availability of faster detectors<sup>34</sup> soon will allow for greatly reduced acquisition times or an increased number of resolved voxels at existing facilities.

**Reconstruction.** The SAXS data were reduced by azimuthal integration within 16 segments. For further analysis we integrated over a  $q$  range of  $0.0379\text{--}0.0758 \text{ nm}^{-1}$ , which corresponds to length scales of 165–85 nm. Projections were aligned with subpixel accuracy on the basis of the absorption measurements on the beam stop diode using an efficient image registration approach based on cross-correlation<sup>35</sup>. A gradient-based optimization algorithm was used to minimize the error metric  $\epsilon = 2 \sum_i \left( \sqrt{\hat{I}_i} - \sqrt{I_i} \right)^2$ , where  $I_i$  is the measured intensity and  $\hat{I}_i$  is the modelled intensity from the  $i$ th scattering pattern. This error metric was chosen because minimizing it is a first-order approximation to a maximum-likelihood estimation for photon-counting Poisson noise<sup>36</sup>. Owing to the high number of parameters to optimize, the gradient of the error metric with respect to such parameters is calculated using an analytical expression rather than finite differences.

The modelled intensity  $\hat{I}$  is obtained by calculating the intensity pattern of each voxel for each orientation and summing them over the corresponding beam

path or ‘projection’. To reduce artefacts due to projections in a rectangular 3D grid along arbitrary orientations, the volume was upsampled by a factor of two before the projection. For the projection, bilinear interpolation was used followed by a triangular 2D filter and by a final downsampling. The reciprocal-space intensity in each voxel was calculated using a series of spherical harmonic functions  $Y_l^m$  of degree  $l$  and order  $m$  as  $\hat{R} = \left| \sum_l \sum_m a_l^m Y_l^m(\theta, \phi) \right|^2$ , where  $\theta$  and  $\phi$  are the polar and azimuthal angles, respectively, and  $a_l^m$  are coefficients of the spherical harmonics. To model the scattering from the collagen fibrils, we used spherical harmonic functions  $Y_l^m$  of degree  $l = 0, 2, 4, 6$  and of order  $m = 0$ . The square of the sum was used to avoid negative intensities. To guide the solution towards the *a priori* known rotational symmetry of the collagen fibrils, smooth penalty terms were introduced for the coefficients  $a_l^m$ . For the gradient calculation, a backprojection operator is needed, which was implemented using bilinear interpolation. To accelerate convergence of the reconstruction, the gradient of the error metric with respect to the coefficients  $a_l^m$  was spatially convolved with a 3D Hamming window of  $3 \times 3 \times 3$  voxels, which can be understood as an application of Grenander’s method of sieves<sup>37</sup>. The parameters of the optimization in each voxel were the polar angle  $\theta_0$  and azimuthal angle  $\phi_0$ , representing the orientation of the probed ultrastructure in 3D as well as the coefficients  $a_l^m$  of the spherical harmonics, from which the degree of orientation  $\sigma$  is calculated as

$$\sigma = \frac{\iint (a_2 Y_2^0 + a_4 Y_4^0 + a_6 Y_6^0)^2 \sin \theta d\phi d\theta}{\iint (a_0 Y_0^0 + a_2 Y_2^0 + a_4 Y_4^0 + a_6 Y_6^0)^2 \sin \theta d\phi d\theta}$$

which corresponds to the ratio between the anisotropic scattering and the total scattering. To accelerate convergence, the optimization was performed following a three-step approach. First, the angles that define the spherical harmonics were optimized while the coefficients were kept fixed, which helps to initially optimize a subset of the parameters. This was followed by optimizing the coefficients of the spherical harmonics while the angles obtained in the first step were kept constant. Typically 20 iterations were performed for these first two steps. Finally, all parameters were optimized simultaneously, with this last step reaching convergence after only two further iterations. Using a development prototype implementation, the reconstruction took three days running on a standard computer server with 12 CPU cores and up to 24 threads process in parallel. The algorithm lends itself well to further parallelization, and an optimized code running on a computer cluster would speed up the processing substantially.

**Code availability.** The computer code that was used for the reconstruction is currently under further development, documentation and optimization and, hence, is currently not publicly accessible. Details of the earlier version of the code as used in the paper are available from the authors on request.

31. Kraft, P. *et al.* Performance of single-photon-counting PILATUS detector modules. *J. Synchrotron Radiat.* **16**, 368–375 (2009).
32. Howells, M. R. *et al.* An assessment of the resolution limitation due to radiation-damage in X-ray diffraction microscopy. *J. Electron Spectrosc.* **170**, 4–12 (2009).
33. Hubbell, J. & Seltzer, M. *Tables of X-ray Mass Attenuation Coefficients and Mass Energy-Absorption Coefficients Version 1.4*. Report No. NISTIR-5632 (National Institute of Standards and Technology, 1995); <http://physics.nist.gov/xaamdi>.
34. Tinti, G., *et al.* Performance of the EIGER single photon counting detector. *J. Instrum.* **10**, C03011 (2015).
35. Guizar-Sicairos, M., Thurman, S. T. & Fienup, J. R. Efficient subpixel image registration algorithms. *Opt. Lett.* **33**, 156–158 (2008).
36. Thibault, P. & Guizar-Sicairos, M. Maximum-likelihood refinement for coherent diffractive imaging. *New J. Phys.* **14**, 063004 (2012).
37. Grenander, U. *Abstract Inference* Ch. 9 (Wiley, 1981).



# Six-dimensional real and reciprocal space small-angle X-ray scattering tomography

Florian Schaffl<sup>1</sup>, Martin Bech<sup>2</sup>, Paul Zaslansky<sup>3</sup>, Christoph Jud<sup>1</sup>, Marianne Liebi<sup>4</sup>, Manuel Guizar-Sicairos<sup>4</sup> & Franz Pfeiffer<sup>1,5</sup>

**When used in combination with raster scanning, small-angle X-ray scattering (SAXS) has proven to be a valuable imaging technique of the nanoscale<sup>1</sup>, for example of bone, teeth and brain matter<sup>2–5</sup>. Although two-dimensional projection imaging has been used to characterize various materials successfully, its three-dimensional extension, SAXS computed tomography, poses substantial challenges, which have yet to be overcome. Previous work<sup>6–11</sup> using SAXS computed tomography was unable to preserve oriented SAXS signals during reconstruction. Here we present a solution to this problem and obtain a complete SAXS computed tomography, which preserves oriented scattering information. By introducing virtual tomography axes, we take advantage of the two-dimensional SAXS information recorded on an area detector and use it to reconstruct the full three-dimensional scattering distribution in reciprocal space for each voxel of the three-dimensional object in real space. The presented method could be of interest for a combined six-dimensional real and reciprocal space characterization of mesoscopic materials with hierarchically structured features with length scales ranging from a few nanometres to a few millimetres—for example, biomaterials such as bone or teeth, or functional materials such as fuel-cell or battery components.**

By raster scanning an object through a focused X-ray beam and recording a scattering pattern at each point, which we refer to as a projection in the following, scanning SAXS imaging combines the ability of SAXS to probe the nanoscale with spatial resolution. Changes in nanostructure over an area many orders of magnitude larger can be revealed using this technique<sup>1</sup>. Because a SAXS pattern is recorded at every scanned point on an area detector (Fig. 1b), it is possible to extract oriented scattering information. This information may be of great importance for the understanding of material properties, and has proven to be very useful for the characterization of highly anisotropic structures, such as collagen<sup>3,12–17</sup>.

Owing to the penetrating nature of X-rays, scanning SAXS may be used to image thick samples. One drawback of pure projection imaging is that the signal is integrated along the beam direction, and consequently all depth information about the scattering structures is lost. It is possible to recover the exact location of a signal in three dimensions using computed tomography. In the case of SAXS, computed tomography has previously been demonstrated, for example, for polymer fibres, amorphous glass and myelin sheaths<sup>6–11</sup>. Compared with other tomographic techniques such as micro computed tomography or ptychographic tomography, which are either unable to go down to nanometre resolution or are limited to very small sample volumes<sup>18,19</sup>, SAXS tomography is able to probe structural information at the nanoscale for objects of comparatively much larger dimensions.

However, only a small fraction of the full two-dimensional (2D) SAXS signal has been able to be used for a correct reconstruction thus far, and, in particular, the orientation information contained in 2D SAXS patterns has not been able to be preserved during tomographic

SAXS reconstruction. We demonstrate the full potential of SAXS tomography, by deriving a way to combine orientation-sensitive scanning SAXS imaging with computed tomography. Therefore we can obtain the full three-dimensional (3D) scattering distribution in each voxel of the 3D space (hence a six-dimensional (6D) combined real and reciprocal space map of the specimen).

For most standard reconstruction techniques, a correct result is only possible if the signal is rotationally invariant, that is, the sum of all line integrals that form a projection image does not depend on the rotation of the sample. The most prominent example where this condition is satisfied is in the attenuation contrast of, for example, neutron or X-ray tomography as used in many industrial and clinical applications. In the case of SAXS tomography, rotational invariance is not generally present in the signal, except in special cases such as when imaging isotropic scattering powders, colloids or samples with certain symmetries. This lack of invariance restricts the number of applications, and excludes a wide range of materials with a preferred scattering orientation, including biomaterials composed of oriented collagen fibres and a wide range of functional materials with fibrous nanostructures.

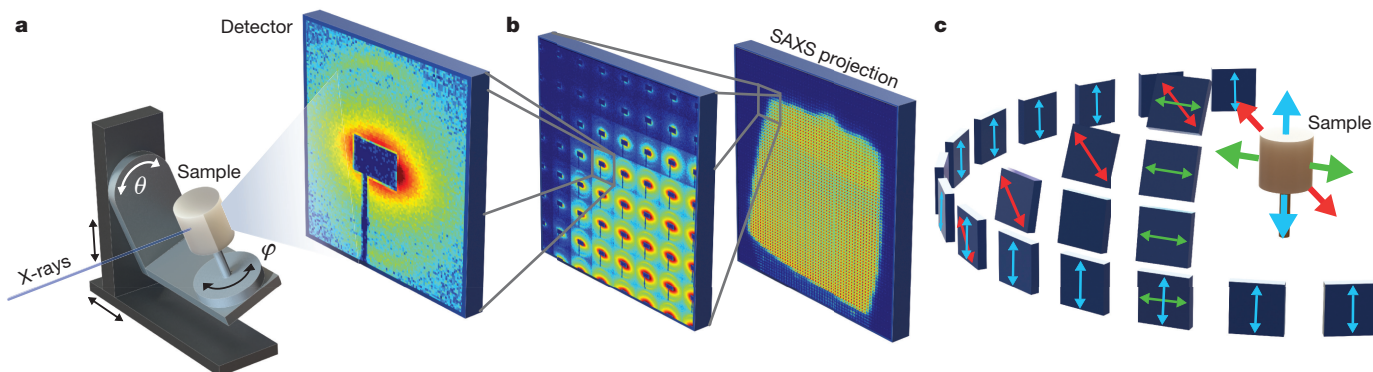
In addition to being observed in isotropic samples, rotational invariance also exists in oriented scattering samples if the momentum transfer vectors  $\mathbf{q}$ , that is, the scattering orientations, are parallel to the rotation axis<sup>9</sup>. This implies that for an arbitrary sample, a correct reconstruction is guaranteed only for  $\mathbf{q}$  vectors parallel to the tomography axis; all other scattering orientations cannot be reconstructed. Therefore, for a complete reconstruction of all possible  $\mathbf{q}$  vectors in three dimensions, additional rotations of the sample are necessary.

One solution to the problem of requiring additional rotations of the sample is to measure and reconstruct the sample from a large number of different rotation axes, each parallel to  $\mathbf{q}$  vectors in one specific direction. This would be highly inefficient and time consuming and thus infeasible in practice. Instead, we add the additional necessary rotations by introducing a tilt to the conventional rotation axis around the horizontal axis perpendicular to the beam direction (Fig. 1a). By doing so, projections are recorded from all possible sample orientations by combining the tomographic rotation  $\varphi$  with a tilt by an angle  $\theta$ . For each of the SAXS patterns of one projection, the scattering vectors  $\mathbf{q}_d$  recorded on the detector represent the integration of 2D slices of the 3D  $\mathbf{q}$  spaces of each voxel along the path of the X-ray beam. For the purpose of tomographic reconstruction, it is thus vital to know which  $\mathbf{q}$  vectors are probed by each projection. This information is obtained by using a projection-dependent rotation matrix  $A_i$  that describes the relationship between the laboratory and sample coordinate systems:

$$\mathbf{q} = A_i^{-1} \mathbf{q}_d \quad (1)$$

If we define the direction of the X-ray beam to be along the  $z$  axis of the laboratory coordinate system, then each projection, indexed  $i$ , is specified in the sample coordinate system by its plane normal vector  $\mathbf{n}_i = A_i^{-1}(0, 0, 1)^T$ .

<sup>1</sup>Lehrstuhl für Biomedizinische Physik, Physik-Department & Institut für Medizintechnik, Technische Universität München, 85748 Garching, Germany. <sup>2</sup>Department of Medical Radiation Physics, Clinical Sciences, Lund, Lund University, 22185 Lund, Sweden. <sup>3</sup>Julius Wolff Institute, Charité - Universitätsmedizin Berlin, 13353 Berlin, Germany. <sup>4</sup>Paul Scherrer Institut, 5232 Villigen PSI, Switzerland. <sup>5</sup>Institut für diagnostische und interventionelle Radiologie, Klinikum rechts der Isar, Technische Universität München, 81675 München, Germany.



**Figure 1 | Experimental set-up and raster-scanning technique.**

**a**, Schematic representation of the experimental set-up. The sample is raster scanned through a focused X-ray beam, indicated by the vertical and horizontal arrows, and a diffraction pattern is recorded at each point. In addition to a rotation  $\varphi$  around the tomography axis, the sample is rotated by  $\theta$  around an additional axis. **b**, Each pixel of a SAXS projection consists of an entire SAXS pattern rather than just a single value. **c**, Schematic

representation of the virtual-tomography-axis technique. Projections recorded at different views of the sample are used for various virtual tomography axes, represented by the red, green and blue arrows within the sample. Each of these axes is reconstructed independently by using appropriate projections and detector segments, as indicated by arrows of matching colour.

Only  $q$  vectors parallel to the tomography axis can be reconstructed, and every SAXS projection probes a 2D slice of the 3D  $q$  space. Consequentially, all projections with orientation  $n_i$  perpendicular to an axis  $t$ , given in sample coordinates, must necessarily contain  $q$  vectors parallel to  $t$  and so fulfil the rotational-invariance requirement for a rotation around  $t$  (see Methods for a more general discussion on rotational invariance).

We now introduce virtual tomography axes. Instead of physically rotating the sample around an axis  $t$ , we select a subset of the full set of available projections that contains all those with  $q$  vectors parallel to the desired axis  $t$ . This technique is illustrated in Fig. 1c. Several projections are sketched in the sample frame of reference. For three select axes  $t$  within the sample, projections with  $q$  vectors parallel to  $t$  are selected, shown in Fig. 1c by arrows of matching colours. The orientation on the detector that contains the rotationally invariant SAXS data for each axis  $t$  and projection is indicated by the respective arrows orientations.

Owing to a finite number of possible recorded projections, it is unlikely that a perfect match between  $t$  and the available projections will be found. We use the scalar product  $|n_i \cdot t|$  to quantify the mismatch between the projection  $n_i$  and the virtual tomography axis  $t$ . For most materials of interest, the SAXS signal generally varies slowly and does not exhibit sharp azimuthal peaks (contrary to crystal diffraction, for example); consequently, projections for which  $0 < |n_i \cdot t| \ll 1$  may be included as well as those for which  $|n_i \cdot t| = 0$ , to ensure a sufficient number of projections for the tomographic reconstruction. Our method is not suitable for materials such as nearly perfect single-crystalline objects, for which the SAXS signal does not vary slowly.

Given such a subset of possible projections, for each individual projection, we need to calculate the orientation of the recorded scattering vectors  $q_d$  corresponding to  $q \parallel t$ . The correct part of the detector is calculated by projecting  $t$  onto the plane specified by  $n_i$ , which allows us to account for the small discrepancy introduced by considering projections for which  $0 \leq |n_i \cdot t| \ll 1$ :  $q_d = t - (t \cdot n_i)n_i$ . Using virtual tomography axes, many different axes  $t$ , and therefore  $q$  vectors, can be reconstructed by selecting the appropriate subsets and corresponding  $q_d$  vectors. A major advantage of this method is that all the information from every SAXS pattern can be used for a complete reconstruction of the 3D  $q$  space in each voxel, rather than just a few  $q_d$  vectors, as is the case when using the standard tomography axis. Removing the requirement of having to directly measure a large number of tomography axes and instead reusing the same projections for many different virtual tomography axes allows us to efficiently perform a real and reciprocal space resolved 6D SAXS computed tomography characterization.

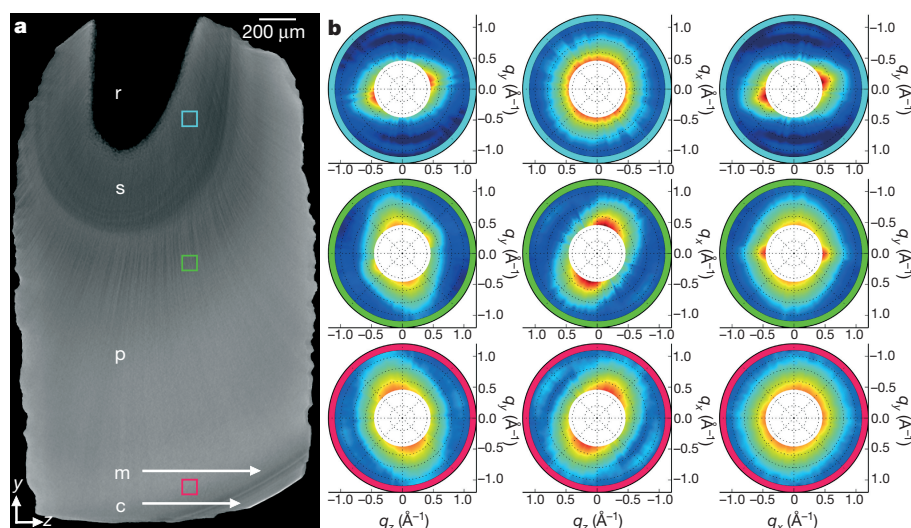
To demonstrate the usefulness of the method, particularly for the characterization of biomaterials, we performed an experiment to

resolve the nanostructures found in a human tooth. Teeth are made of a complex, highly hierarchically structured material that is able to withstand years of cyclic load in the harsh environment of the mouth. Knowledge of the exact structural details of this material at length scales spanning the nanoscale to the macroscale is still missing, and would help to understand the basic structure–function relationships that have been optimized during millions of years of evolution. This understanding is important for the design of tailored synthetic materials that mimic the durability and stiffness of teeth. An overview of four different tissues present in the investigated sample is shown in Fig. 2a. Mean mineral density information, observed along the long axis of the sample and obtained using high-resolution, state-of-the-art absorption tomography, reveals the layers of cementum and of mantle, primary and secondary dentine in the sample. All these tissues are composed of mineralized collagen fibres that give rise to a strong characteristic SAXS signal, owing to the periodic arrangement of mineral platelets along the fibre axis with approximately 67.6-nm spacing<sup>3</sup>. However, the details of the nano-architecture that is essential for the mechanical performance remains unresolved.

Using the SAXS computed tomography technique described above, we reconstructed the full 3D scattering distribution in every voxel for the whole  $q$  range, rather than reconstructing just one value as in conventional tomography. From this data, we virtually extract information about what the reciprocal scattering space from each voxel looks like from any given orientation. Planar cuts along the coordinate-system axes of this voxel-wise 3D information are shown in Fig. 2b for three different regions within the sample. The top, blue-ringed slices of Fig. 2b corresponding to the pulp region of the tooth sample, which consists of secondary dentine, have a pronounced vertical fibre orientation, seen as clear collagen peaks in the  $q_y$  direction. Different orientations and scattering signal strengths are seen in the green- and red-ringed slices, which correspond to regions of the tooth that are composed of primary and mantle dentine.

Given the abundance of reconstructed data, a major difficulty is to extract the information of interest and compress the results into an easily understandable form. One possibility for determining collagen orientation is as follows. As well as diffuse background scattering that stems from the fibre bundles, collagen fibres display distinct peaks around  $|q| \approx 0.9 \text{ \AA}^{-1}$ . Previous studies<sup>20</sup> have subtracted the background scattering and isolated the collagen peak by fitting a power-law function with an exponent of about 2.6 (ref.). To extract the orientation of the collagen peaks, we took the ratio of  $q$ -vector intensities in the range with  $(0.88\text{--}0.94 \text{ \AA}^{-1})$  and without  $(0.82\text{--}0.86 \text{ \AA}^{-1})$  collagen peaks. Assuming a constant exponent for the background scattering, the ratio of  $q$ -vector intensities deviates from a constant value only





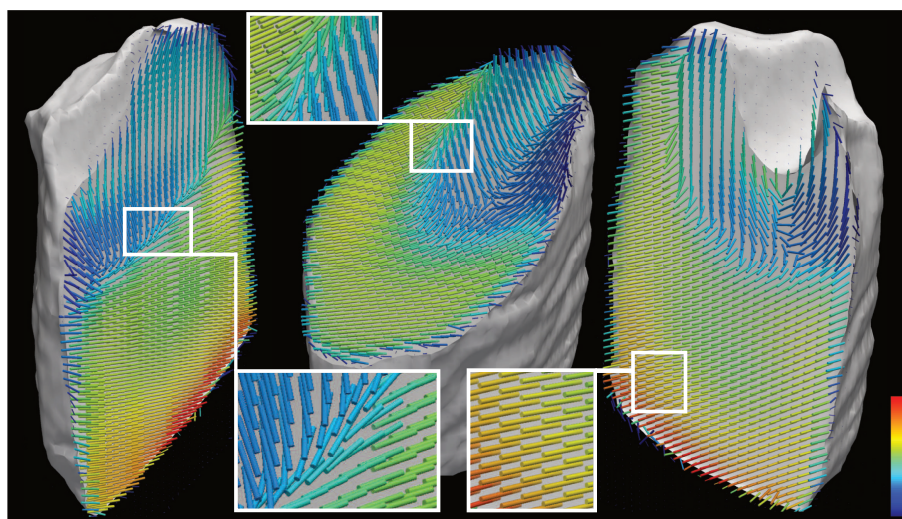
**Figure 2 | Virtual SAXS patterns extracted from the reconstructed 3D volume.** **a**, Average density within the sample obtained using state-of-the-art absorption micro computed tomography. Distinct dental tissues found in the sample are identified around the root canal (r): secondary dentine (s), primary dentine (p), mantle dentine (m) and cementum (c). **b**, 2D slices along the coordinate axes of the reconstructed 3D  $\mathbf{q} = (q_x, q_y, q_z)$  space. The logarithmic scattering strength for three points in

the centre of the sample, indicated by the blue, green and red squares, is shown (colour coded by the outer ring on the plots) for slices in the  $q_z$ - $q_y$ ,  $q_z$ - $q_x$  and  $q_x$ - $q_y$  planes. Distinct collagen peaks can be seen at  $|q| \approx 0.9 \text{ \AA}^{-1}$ , except for slices roughly perpendicular to the fibre orientation. The orientation of the fibres and the intensity of the peaks change noticeably in the different regions of the sample. An animation showing multiple slices is provided in Supplementary Video 1.

in those directions with additional scattering caused by the distinct periodicity of the collagen fibres. From the resulting 3D distribution of ratios, we extracted the mean fibre orientation by ellipsoid fitting and determining the largest principal axis. The resultant orientations for example slices inside the sample are shown in Fig. 3. The fibre orientations are represented by small bars whose colour represents the average scattering intensity in the collagen range  $0.88\text{--}0.94 \text{ \AA}^{-1}$ . Their length is scaled with the respective absorption values, so that equally long and vanishingly small bars are obtained inside and outside the sample volume, respectively. A strong change in collagen fibre orientation near the root-canal region (upper side of Fig. 3) of the sample contrasts the gradual increase in scattering strength towards the outer sides of the root (lower side of Fig. 3). The collagen fibre orientation is just one example of many possible parameters that can be extracted from the 6D data. Various other parameters have been extracted from SAXS data previously, such as the local degree of alignment of the mineral

platelets contained within the collagen matrix, or their shape and size distribution<sup>21</sup>.

In conclusion, we demonstrated a combined 6D real and reciprocal space SAXS computed tomography technique as a method to simultaneously obtain information about the structural features of a sample from the nanometre to the millimetre scale. Using virtual rotation axes, we take advantage of all data recorded and are able to reconstruct the full 3D scattering distribution in each voxel of the 3D real space of the sample. This vast amount of additional information in each voxel could enable analysis of the 3D nanostructure of materials, spatially resolved in 3D on a large scale. As with all imaging techniques, analysis of the resultant data is a big part of the work. Given the amount of data reconstructed by the presented method, there is a need for data analysis tailored to a specific problem. For the purpose of the investigated collagen sample, we proposed one possible way to extract fibre orientations in three dimensions. Although our approach is a substantial step



**Figure 3 | 3D visualization of collagen fibre orientation within the tooth sample.** The orientation of the coloured bars indicate the mean orientation of collagen fibres obtained from ellipsoid fits to the ratios of scattering intensities of  $\mathbf{q}$  ranges with and without collagen peaks. The

colour represents the average scattered intensity in the collagen range  $0.88\text{--}0.94 \text{ \AA}^{-1}$ . The underlying 3D nanomorphology within the entire sample is revealed. An animation showing all slices in one direction is provided in Supplementary Video 2.



forward, measurement and reconstruction times were considerable. (A discussion of the prospects of advances in hardware and the subsequent reduction in time needed for an experiment is given in Methods.) Nonetheless, we believe that a combined 6D real and reciprocal space characterization is of great interest for nanostructure characterization of mesoscopic materials and composites with hierarchically structured features ranging from the few-nanometre to the few-millimetre length scale. Possible applications of such a characterization include the study of natural bones and teeth, or man-made functional materials.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 25 June; accepted 22 September 2015.**

1. Fratzl, P., Jakob, H. F., Rinnerthaler, S., Roschger, P. & Klaushofer, K. Position-resolved small-angle X-ray scattering of complex biological materials. *J. Appl. Cryst.* **30**, 765–769 (1997).
2. Fratzl, P., Schreiber, S. & Boyde, A. Characterization of bone mineral crystals in horse radius by small-angle X-ray scattering. *Calcif. Tissue Int.* **58**, 341–346 (1996).
3. Kinney, J. H., Pople, J. A., Marshall, G. W. & Marshall, S. J. Collagen orientation and crystallite size in human dentin: a small angle X-ray scattering study. *Calcif. Tissue Int.* **69**, 31–37 (2001).
4. Falzon, G. *et al.* Myelin structure is a key difference in the x-ray scattering signature between meningioma, schwannoma and glioblastoma multiforme. *Phys. Med. Biol.* **52**, 6543–6553 (2007).
5. De Felici, M. *et al.* Structural characterization of the human cerebral myelin sheath by small angle x-ray scattering. *Phys. Med. Biol.* **53**, 5675–5688 (2008).
6. Schroer, C. G. *et al.* Mapping the local nanostructure inside a specimen by tomographic small-angle x-ray scattering. *Appl. Phys. Lett.* **88**, 164102 (2006).
7. Stribeck, N. *et al.* Volume-resolved nanostructure survey of a polymer part by means of SAXS microtomography. *Macromol. Chem. Phys.* **207**, 1139–1149 (2006).
8. Stribeck, N. *et al.* SAXS-fiber computer tomography. Method enhancement and analysis of microfibrillar-reinforced composite precursors from PEBA and PET. *Macromolecules* **41**, 7637–7647 (2008).
9. Feldkamp, J. M. *et al.* Recent developments in tomographic small-angle X-ray scattering. *Phys. Status Solidi A* **206**, 1723–1726 (2009).
10. Jensen, T. H. *et al.* Brain tumor imaging using small-angle x-ray scattering tomography. *Phys. Med. Biol.* **56**, 1717–1726 (2011).
11. Jensen, T. H. *et al.* Molecular X-ray computed tomography of myelin in a rat brain. *Neuroimage* **57**, 124–129 (2011).
12. Rinnerthaler, S. *et al.* Scanning small angle X-ray scattering analysis of human bone sections. *Calcif. Tissue Int.* **64**, 422–429 (1999).
13. Boote, C., Dennis, S. & Meek, K. Spatial mapping of collagen fibril organisation in primate cornea—an X-ray diffraction investigation. *J. Struct. Biol.* **146**, 359–367 (2004).
14. Moger, C. J. *et al.* Regional variations of collagen orientation in normal and diseased articular cartilage and subchondral bone determined using small angle X-ray scattering (SAXS). *Osteoarthr. Cartilage* **15**, 682–687 (2007).
15. Bunk, O. *et al.* Multimodal x-ray scatter imaging. *New J. Phys.* **11**, 123016 (2009).
16. Meek, K. M. & Boote, C. The use of X-ray scattering techniques to quantify the orientation and distribution of collagen in the corneal stroma. *Prog. Retin. Eye Res.* **28**, 369–392 (2009).
17. Märten, A., Fratzl, P., Paris, O. & Zaslansky, P. On the mineral in collagen of human crown dentine. *Biomaterials* **31**, 5479–5490 (2010).
18. Bossa, N. *et al.* Micro- and nano-X-ray computed-tomography: a step forward in the characterization of the pore network of a leached cement paste. *Cement Concr. Res.* **67**, 138–147 (2015).
19. Dierolf, M. *et al.* Ptychographic X-ray computed tomography at the nanoscale. *Nature* **467**, 436–439 (2010).
20. Gaiser, S., Deyhle, H., Bunk, O., White, S. N. & Müller, B. Understanding nano-anatomy of healthy and carious human teeth: a prerequisite for nanodentistry. *Biointerphases* **7**, 4 (2012).
21. Fratzl, P., Gupta, H., Paschalis, E. P. & Roschger, P. Structure and mechanical quality of the collagen-mineral nano-composite in bone. *J. Mater. Chem.* **14**, 2115–2123 (2004).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** The SAXS experiments were performed at the cSAXS beamline of the Swiss Light Source (SLS) at the Paul Scherrer Institut (PSI), Villigen, Switzerland. We are grateful for travel support that was granted by the EU access program CALIPSO. We acknowledge financial support through the DFG Cluster of Excellence Munich-Centre for Advanced Photonics (MAP) and the DFG Gottfried Wilhelm Leibniz program. P.Z. is grateful for funding of the DFG (German Research Foundation) through SPP1420. F.S. and C.J. thank the TUM Graduate School for support of their studies. F.P. acknowledges support through the TUM Institute for Advanced Studies (TUM-IAS). We thank A. Fehrer for developing the GPU projectors used for the reconstruction.

**Author Contributions** F.P., M.B. and F.S. conceived the experiment. M.L. and M.G.-S. developed the sample stage used in the experiment. M.B., C.J., M.L., M.G.-S. and F.S. performed the experiment at cSAXS/PSI. P.Z. prepared the tooth sample. Data analysis was performed by F.S. with input and discussion from M.B., F.P. and P.Z. F.S. wrote the manuscript with contributions from all co-authors.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to F.S. ([florian.schaff@tum.de](mailto:florian.schaff@tum.de)), M.B. ([martin.bech@med.lu.se](mailto:martin.bech@med.lu.se)) or F.P. ([franz.pfeiffer@tum.de](mailto:franz.pfeiffer@tum.de)).

## METHODS

**Sample.** A human upper lateral incisor was obtained from a pool of teeth collected from patients undergoing routine dental treatment unrelated to this study. The tooth was obtained with written informed consent, following the directives of the Ethical Review Committee of the Charité - Universitätsmedizin Berlin (EA4/002/09). The tooth had no caries and was extracted for periodontal reasons from an anonymous donor. For the SAXS experiment, a cylindrical segment of the tooth about 3 mm in diameter and 4 mm in length was extracted using water-cooled dental drills. The harvested piece of the root was extracted from a region beneath the tooth crown, facing the lip (buccal side of the tooth).

**Experiment.** The measurements were conducted at the X12SA (cSAXS) beamline of the Swiss Light Source (Paul-Scherrer Institute) with monochromatic 18.6-keV X-rays. A pencil beam of  $50 \mu\text{m} \times 50 \mu\text{m}$  at the sample position was used to raster scan the specimen with a matching step size of  $50 \mu\text{m}$  and exposure time of 50 ms per point. In total, scanning SAXS projections were recorded from 288 unique sample orientations. The 288 projections were recorded for 10 different tilt angles  $\theta = 0^\circ, -4^\circ, -12^\circ, -20^\circ, -28^\circ, -36^\circ, -44^\circ, -52^\circ, -60^\circ, -68^\circ$  of the tomography axis. A varying number (55, 29, 29, 29, 29, 29, 29, 25, 19, 15) of tomography angles  $\varphi$ , evenly distributed over  $360^\circ$ , was used for each of the tilted tomography scans. The  $\theta = -12^\circ, -28^\circ, -44^\circ$  and  $-60^\circ$  tomographic measurements included a  $\varphi$  offset of  $6.2^\circ$  so a more even distribution of the projections could be achieved. Each projection consisted of  $59 \times 81$  SAXS patterns, in the horizontal and vertical directions, respectively, whereby the horizontal axis was scanned in continuous mode. In total, 1,376,352 SAXS patterns were recorded in slightly less than 40 h. The scattering signal was collected using a photon-counting PILATUS 2M detector<sup>22</sup> placed 7,363 mm downstream of the sample. Simultaneously, X-ray attenuation data was acquired using a diode directly mounted on the beamstop, which was used for a necessary normalization of the scattering signal<sup>10</sup> and registration of the projections. The normalization was done by dividing each recorded SAXS pattern at point  $p$  (including all variables describing sample position and rotation) by the relative absorption of the sample  $I^{\text{diode}}(p)/I_0^{\text{diode}}$ :

$$I_{\text{norm}}^{\text{SAXS}}(q, p) = I^{\text{SAXS}}(q, p) \frac{I_0^{\text{diode}}}{I^{\text{diode}}(p)}$$

**Reconstruction.** Azimuthal integration was performed using pyFAI<sup>23</sup> for all recorded SAXS patterns, regrouping the data into 360 azimuthal segments with an extent of  $1^\circ$  each. Following the integration, the data were reduced to an azimuthal range of  $180^\circ$  by averaging, making use of the fact that the data are symmetric with respect to the centre of the diffraction pattern.

The selection of projections for each virtual tomography axis included an allowed deviation of the scalar product  $|\mathbf{n}_i \cdot \mathbf{t}| < 0.05$  throughout the reconstruction. For our 288 recorded projections, this condition ensures that a sufficient number of points are picked, without increasing the introduced error more than necessary.

A virtual tomography axis is expressed in sample coordinates by  $\mathbf{t} = A(\theta, \varphi)^{-1}(0, 0, 1)^T$ . The reconstructed virtual tomography axes were chosen for 30 different values of  $\theta$  from  $0^\circ$  to  $90^\circ$  and a varying number  $N_\varphi = 60\sin(\theta\pi/180) + 1$  of values of  $\varphi$  from  $0^\circ$  to  $360^\circ$ , depending on the current value of  $\theta$ . This choice of  $N_\varphi$  ensured that the virtual tomography axes were chosen to cover a solid angle of  $2\pi$ , allowing for a complete reconstruction of all possible orientations.

In total, 1,168 evenly distributed virtual tomography axes in 13 different  $q$ -vector intervals ranging from  $|q| = 0.45 \text{ \AA}^{-1}$  to  $|q| = 1.16 \text{ \AA}^{-1}$  were reconstructed; 15,184 individual reconstructions were performed. Each of these consists of a unique set of projections and azimuthal segments. Because only a very small azimuthal part of the detector is used from each SAXS pattern, the radial width of the intervals was chosen to integrate over several pixels to increase photon statistics. The closest azimuthal segment to the projected virtual tomography axis was taken, that is, an azimuthal error of less than  $0.5^\circ$ . Each of the reconstructions was performed using a SART (simultaneous algebraic reconstruction technique) algorithm with total variation regularization, because this reconstruction technique has been shown<sup>24</sup> to perform well with strong undersampling and missing angular wedges, as in the present case. The computing time per reconstruction was about 40 s on a computer built for GPU computing ( $2 \times$  Intel Xeon E5-2643,  $4 \times$  Nvidia Tesla Kepler K10, 256 GB RAM). Therefore, the reconstruction time for all reconstructions amounts to slightly over one week. In principle, all reconstructions can be run in parallel because they are completely independent from each other. An optimized, parallel reconstruction would thus allow for a substantial reduction in computing time—potentially to a few hours.

To verify a correct reconstruction result, intensity maps of the reconstructed and reprojected scattering vectors alongside their measured counterparts are shown in Extended Data Fig. 1 for the  $|q| = 0.88\text{--}0.94 \text{ \AA}^{-1}$  interval at two select azimuthal scattering orientations of one projection.

The 3D visualization of the reconstruction results was done using ParaView (<http://www.paraview.org>).

**Absorption computed tomography.** The high-resolution micro computed tomography of the sample was acquired with a VersaXRM-500 X-ray microscope, Xradia (Pleasanton), operated at 50 kVp, 4 W and 30 s exposure time. A voxel size of  $3.087 \mu\text{m}$  was achieved.

**Coordinate system.** To describe the necessary rotations it is important to define coordinate systems, which was done according Extended Data Fig. 2. A fixed laboratory coordinate system is given by  $x_{\text{lab}}, y_{\text{lab}}, z_{\text{lab}}$ , with  $z_{\text{lab}}$  in the beam direction,  $y_{\text{lab}}$  pointing upwards and  $x_{\text{lab}}$  completing the right-handed coordinate system. During the acquisition of each projection, the sample is raster scanned along the  $x_{\text{lab}}$  and  $y_{\text{lab}}$  axes for a total of  $n \times m$  SAXS patterns per projection. For the reconstruction, a sample coordinate system is defined by its axes  $x, y, z$ , which are equivalent to the laboratory coordinate axes prior to any rotation. The rotations required for 3D SAXS computed tomography are described using Euler angles  $\psi, \theta$  and  $\varphi$ , which follow a YZY convention. For our purpose, it is convenient to set  $\psi = 90^\circ$  so that  $\varphi$  and  $\theta$  represent the tomographic rotation and its tilt around  $x_{\text{lab}}$ . For all sample rotations, the corresponding rotation matrix  $A_i$  is calculated as a combination of three successive rotations:

$$\begin{aligned} A_i &= A_\psi A_\theta A_\varphi \\ &= \begin{pmatrix} \cos\psi & 0 & \sin\psi \\ 0 & 1 & 0 \\ -\sin\psi & 0 & \cos\psi \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cos\varphi & 0 & \sin\varphi \\ 0 & 1 & 0 \\ -\sin\varphi & 0 & \cos\varphi \end{pmatrix} \\ &= \begin{pmatrix} -\sin\psi/\sin\varphi + \cos\theta\cos\psi/\cos\varphi & -\sin\theta\cos\psi & \sin\psi\cos\varphi + \cos\theta\cos\psi\sin\varphi \\ \sin\theta\cos\varphi & \cos\theta & \sin\theta\sin\varphi \\ -\cos\psi/\sin\varphi - \cos\theta\sin\psi/\cos\varphi & \sin\theta\sin\psi & \cos\psi\cos\varphi - \cos\theta\sin\psi\sin\varphi \end{pmatrix} \end{aligned}$$

The scattering vectors  $\mathbf{q}_d = (d_x, d_y, 0)^T$  recorded on the detector are described using detector coordinates  $d_x$  and  $d_y$ . The relationship between  $\mathbf{q}_d$  and  $\mathbf{q}$  for each projection is given according to equation (1) as  $\mathbf{q} = A_i^{-1}\mathbf{q}_d$ .

**Rotational invariance.** Following ref. 25, a rotational invariance check is performed for a standard, vertical tomography axis as follows. For any horizontal line  $y_{\text{lab}}$  of a projection  $P$ , a value  $I_{P, y_{\text{lab}}}(d_x, d_y) = \sum_{x=0}^n I_{P, y_{\text{lab}}}(d_x, d_y, x_{\text{lab}})$  is assigned to each pixel of the detector as the sum of all pixel values for this specific line scan. This function necessarily must be constant under rotation if rotational invariance is given—the information changes its horizontal position, but is not lost. Pixel-wise rotational invariance is then calculated as the ratio of the standard deviation to the mean of  $I_{P, y_{\text{lab}}}(d_x, d_y)$  over all projections  $P$ :

$$\frac{\sigma(I_{P, y_{\text{lab}}}(d_x, d_y))}{I_{P, y_{\text{lab}}}(d_x, d_y)}$$

A result for the vertical tomography axis of our measurement is shown in Extended Data Fig. 3. The pixel-wise rotational invariance is shown for two different slices of the sample, marked by the white lines. Because a rotation is only performed around the vertical axis, these slices are independent from each other. Whereas the upper slice (Extended Data Fig. 3b) displays certain rotational invariance for pixels not on the vertical axis, owing to fibre symmetry, the lower slice (Extended Data Fig. 3c) clearly shows that rotational invariance is given only in the vertical orientation.

We want to extend the rotational-invariance check to more than one rotation axis, but a line-wise treatment has proven to be insufficient. Because of additional rotations, different slices of the sample are no longer independent from each other, and the sample needs to be treated as a whole. Rather than summing over one scanning line, we have to sum over all scattering patterns obtained for one projection:  $I_P(d_x, d_y) = \sum_{x_{\text{lab}}, y_{\text{lab}}=0}^{n, m} I_P(d_x, d_y, x_{\text{lab}}, y_{\text{lab}})$ . Additionally, for an arbitrary virtual rotation axis  $\mathbf{t}$ , projections from a tilted rotation axis are included in the reconstruction. From this it follows that the rotationally invariant data are no longer restricted to the vertical direction, but are found in different azimuthal segments on the detector for each projection  $P$ . Thus, the scattering patterns need to be rotated accordingly for a proper comparison. This rotation is most easily done with the already azimuthally regrouped data  $I_P(\chi, r)$ , with  $\chi$  the azimuthal and  $r$  the radial coordinate. A rotation of the data then simplifies to a projection-dependent offset of  $\chi$ , and rotational invariance can be checked for any arbitrary set of projections and virtual tomography axis  $\mathbf{t}$ . Extended Data Figure 4 shows rotational-invariance results for two different virtual tomography axes— $\mathbf{t} = (0, 1/\sqrt{2}, 1/\sqrt{2})$  (Extended Data Fig. 4a) and  $\mathbf{t} = (1/\sqrt{2}, 1/\sqrt{2}, 0)$  (Extended Data Fig. 4b)—for already azimuthally regrouped data. For each pattern, all data has been shifted so that  $\chi = 0^\circ$  or  $\chi = 180^\circ$  corresponds to the detector

orientation parallel to the chosen virtual rotation axis. Rotational invariance is achieved for any given virtual tomography axis by using data only from projections and orientations with scattering parallel to this axis.

**Discussion on acquisition times and resolution.** The number of SAXS patterns that can be collected in a given time frame is mostly determined by two key parameters of the X-ray beam used in the experiment—the X-ray energy and flux of the beam. To reduce the exposure time needed per single SAXS pattern, it is necessary to have as many photons contribute to the SAXS signal as possible. We measured the tooth sample presented in this work at 18.6 keV, the highest available photon energy at the cSAXS beamline. At this energy, a large proportion of the X-rays were absorbed in our sample and therefore did not contribute to the SAXS signal. Even a small increase in photon energy would have a very noticeable effect on the transmission, and ultimately the exposure time required for each point. The second major factor for reduction of the scanning time needed is the flux available for the experiment. Even though synchrotron facilities today are already very powerful, their technology is continuously being improved. An example is the upgrade being installed at the European Synchrotron Radiation Facility (ESRF). Planned to be finished by 2018, this upgrade is expected to increase the available photon flux by up to two orders of magnitude. These ongoing developments will enable very rapid data acquisition to be performed in the future, facilitated by recently developed fast-framing pixel detectors with frame rates of several kilohertz<sup>26</sup>.

A point of concern arising from much stronger sources is the issue of radiation damage. During our experiment we performed several control-scans throughout the measurement and confirmed that there was no change in the signal even after several hours of beam exposure. It is hard to predict at what point radiation damage inflicted by the beam starts to become an issue. There are, however, possibilities to limit these effects, such as active cooling with cryo-jets.

The achievable real-space resolution in SAXS computed tomography is dictated mainly by the investigated sample and number of SAXS patterns (points) available. The total number of projections strongly depends on the number of individual scanning points per projection. From a tomographic point of view, there is little benefit from measuring only a few projections with a large number of points each or, similarly, a large number of projections with only a few points each. Therefore, the number of points per projection cannot be scaled up easily without a massive increase in the acquisition time required. As a consequence, the real-space resolution is strongly limited by the size of the sample. For reference, the real-space

resolution of our measurement for a 4-mm object was about 100  $\mu\text{m}$ , because of the continuous acquisition used.

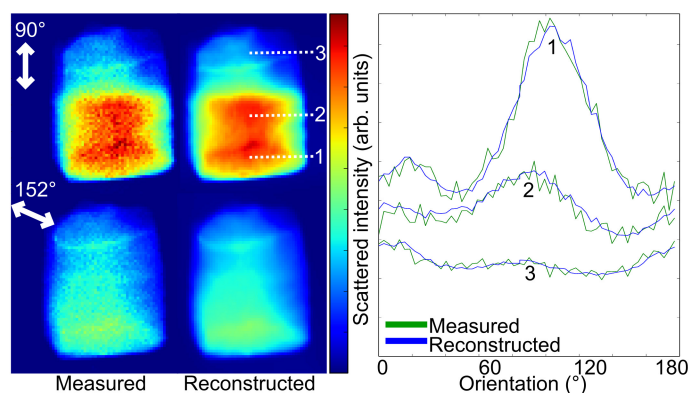
The resolution in reciprocal space has to be discussed for two different cases, namely radial and angular resolution in reciprocal space. The radial resolution is mostly limited by the angular divergence of the beam and the pixel size of the detector. We chose to radially integrate over several detector pixels, which limits our radial reciprocal-space resolution to  $0.05 \text{ \AA}^{-1}$ . In contrast to this, the angular resolution in reciprocal space is largely limited by the number of projections recorded. This number determines the deviation allowed when selecting projections for each virtual tomography axis. In our case, a deviation of the scalar product  $|\mathbf{n}_i \cdot \mathbf{t}| < 0.05$  corresponds to a minimal angular resolution of slightly less than  $6^\circ$ .

Considering all these points, the greatest potential to improve the performance of our method lies in the use of stronger sources and higher photon energy. Our measurement took 40 h. An increase in usable flux by a factor of ten would decrease the time needed for exactly the same measurement by approximately the same factor—to slightly less than 4 h. This reduction would enable SAXS computed tomography to be used for a case study of several samples. A substantial reduction in acquisition time can also be achieved by scanning projections with lower real-space and angular reciprocal-space resolution, owing to the relationships mentioned previously.

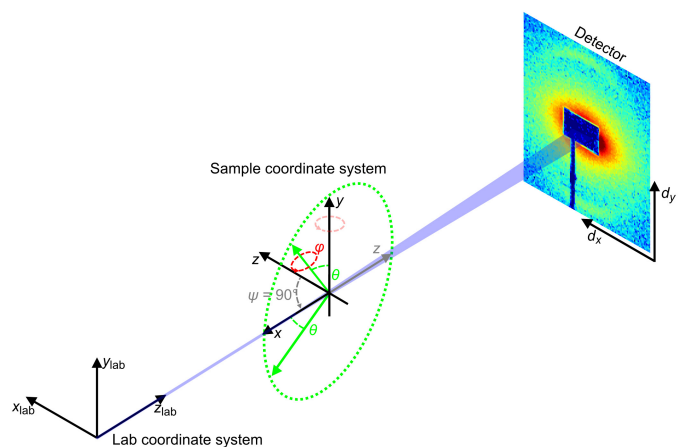
**Code availability.** The code used for azimuthal regrouping is openly available at <https://github.com/kif/pyFAI>. Additional code used is available upon request.

22. Eikenberry, E. F. *et al.* PILATUS: a two-dimensional X-ray detector for macromolecular crystallography. *Nucl. Instrum. Methods Phys. Res. A* **501**, 260–266 (2003).
23. Kieffer, J. & Karkoulis, D. PyFAI, a versatile library for azimuthal regrouping. *J. Phys. Conf. Ser.* **425**, 202012 (2013).
24. Sidky, E. Y., Kao, C.-M. & Pan, X. Accurate image reconstruction from few-views and limited-angle data in divergent-beam CT. *J. X-ray. Sci. Tech.* **14**, 119–139 (2006).
25. Feldkamp, J. *Scanning Small-angle X-ray Scattering Tomography. Non-destructive Access to the Local Nanostructure*. PhD thesis, Technische Universität Dresden (2009); <http://nbn-resolving.de/urn:nbn:de:bsz:14-qucosa-24925>.
26. Dinapoli, R. *et al.* ELGER: next generation single photon counting detector for X-ray applications. *Nucl. Instrum. Methods Phys. Res. A* **650**, 79–83 (2011).



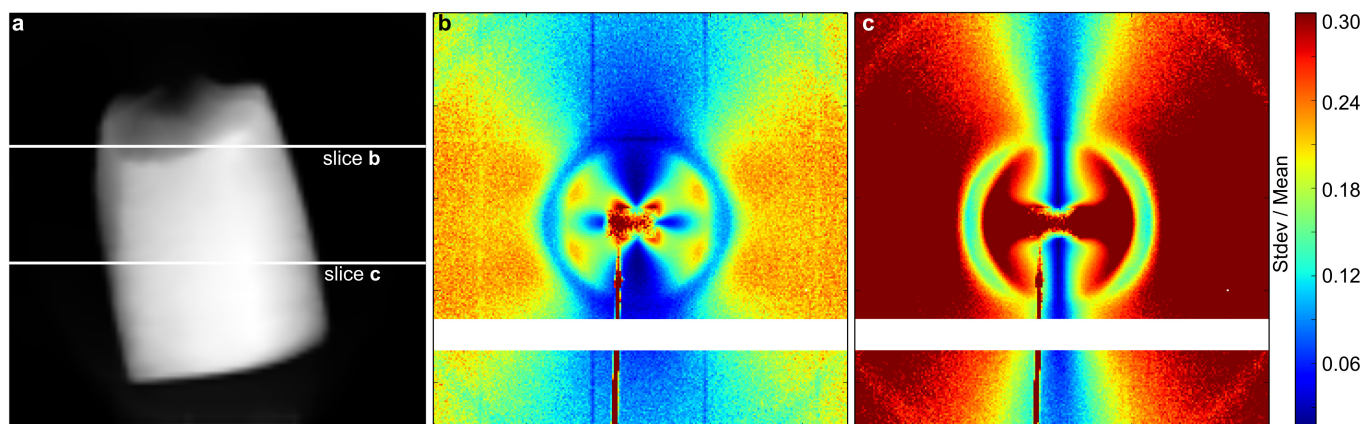


**Extended Data Figure 1 | Comparison of measured and reconstructed values for one projection.** Left, reprojected  $q$  data for two different orientations ( $90^\circ$  and  $152^\circ$ ) is compared to the measured data. Right, azimuthal values for both the reconstructed and measured data are given for three select points, indicated by the dashed lines in the left panel. For the chosen  $|q|$  range, distinct collagen peaks are reconstructed correctly for points 1, 2 and 3 at around  $15^\circ$ ,  $0^\circ$  and  $90^\circ$ , respectively. A good agreement between reconstruction and measurement is seen. Animations of this and further projections showing all  $q$  orientations are provided in Supplementary Videos 3–5.



### Extended Data Figure 2 | Coordinate system used for the experiment.

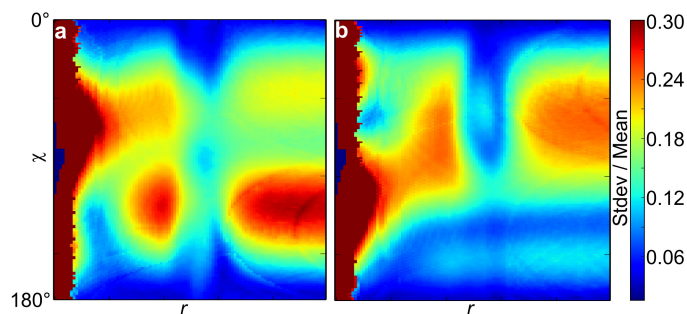
The sample orientation is described using three Euler angles  $\theta$ ,  $\varphi$  and  $\psi$ . With  $\psi = 90^\circ$ ,  $\theta$  represents a tilt of the tomography axis around  $x_{lab}$  and  $\varphi$  describes a rotation around this tilted axis. The sample is scanned along  $x_{lab}$  and  $y_{lab}$ , with a diffraction pattern collected at each point. The detector coordinates are given by  $d_x$  and  $d_y$ .



**Extended Data Figure 3 | Rotational invariance for a standard SAXS computed tomography with a vertical tomography axis.** **a**, Absorption image acquired from the diode data with two vertical slices marked. **b**, **c**, Rotational invariance as defined in the text for both slices. In both cases, rotational invariance is present for all pixels that correspond to scattering orientations parallel to the vertical rotation axis. The collagen

fibres in the top part of the sample, shown in **b**, are mainly vertical. Owing to this symmetry, rotational invariance is also present for pixels not on the vertical axis. Without this symmetry, rotational invariance exists only for the vertical direction, as seen in **c**. The white bars in the lower half of the images are areas between the individual detector modules.





**Extended Data Figure 4 | Generalized rotational invariance.**

**a, b,** Rotational invariance shown for different virtual tomography axes:  $\mathbf{t} = (0, 1/\sqrt{2}, 1/\sqrt{2})$  (**a**) and  $\mathbf{t} = (1/\sqrt{2}, 1/\sqrt{2}, 0)$  (**b**). Radially integrated data are used. Compared to the standard case, shown in Extended Data Fig. 3, a line-wise integration is not possible in the general case. Instead, the standard deviation and mean are calculated from the projection-wise sum of all integrated SAXS patterns. A shift of the azimuthal angle so that the scattering orientation parallel to  $\mathbf{t}$  is at  $0^\circ$  was applied. As can be seen, rotational invariance is also achieved in the general case for scattering orientations parallel to  $\mathbf{t}$ .

# Methane storage in flexible metal–organic frameworks with intrinsic thermal management

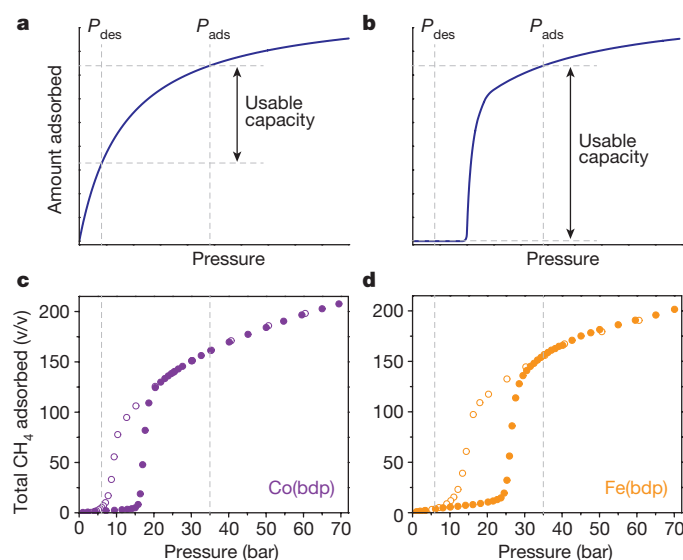
Jarad A. Mason<sup>1</sup>, Julia Oktawiec<sup>1</sup>, Mercedes K. Taylor<sup>1</sup>, Matthew R. Hudson<sup>2</sup>, Julien Rodriguez<sup>3</sup>, Jonathan E. Bachman<sup>1</sup>, Miguel I. Gonzalez<sup>1</sup>, Antonio Cervellino<sup>4</sup>, Antonietta Guagliardi<sup>5</sup>, Craig M. Brown<sup>2,6</sup>, Philip L. Llewellyn<sup>3</sup>, Norberto Masciocchi<sup>7</sup> & Jeffrey R. Long<sup>1</sup>

As a cleaner, cheaper, and more globally evenly distributed fuel, natural gas has considerable environmental, economic, and political advantages over petroleum as a source of energy for the transportation sector<sup>1,2</sup>. Despite these benefits, its low volumetric energy density at ambient temperature and pressure presents substantial challenges, particularly for light-duty vehicles with little space available for on-board fuel storage<sup>3</sup>. Adsorbed natural gas systems have the potential to store high densities of methane (CH<sub>4</sub>, the principal component of natural gas) within a porous material at ambient temperature and moderate pressures<sup>4</sup>. Although activated carbons, zeolites, and metal–organic frameworks have been investigated extensively for CH<sub>4</sub> storage<sup>5–8</sup>, there are practical challenges involved in designing systems with high capacities and in managing the thermal fluctuations associated with adsorbing and desorbing gas from the adsorbent. Here, we use a reversible phase transition in a metal–organic framework to maximize the deliverable capacity of CH<sub>4</sub> while also providing internal heat management during adsorption and desorption. In particular, the flexible compounds Fe(bdp) and Co(bdp) (bdp<sup>2–</sup> = 1,4-benzenedipyrazolate) are shown to undergo a structural phase transition in response to specific CH<sub>4</sub> pressures, resulting in adsorption and desorption isotherms that feature a sharp ‘step’. Such behaviour enables greater storage capacities than have been achieved for classical adsorbents<sup>9</sup>, while also reducing the amount of heat released during adsorption and the impact of cooling during desorption. The pressure and energy associated with the phase transition can be tuned either chemically or by application of mechanical pressure.

The driving range of an adsorbed natural gas (ANG) vehicle is determined primarily by the volumetric usable CH<sub>4</sub> capacity of the adsorbent, which is defined as the difference between the amount of CH<sub>4</sub> adsorbed at the target storage pressure (generally 35–65 bar) and the amount that is still adsorbed at the lowest desorption pressure (generally 5.8 bar)<sup>8–10</sup>. With few exceptions<sup>11</sup>, adsorbents that have been investigated in the context of natural gas storage exhibit classical Langmuir-type adsorption isotherms, where the amount of CH<sub>4</sub> adsorbed increases continuously, but at a decreasing rate, as the pressure is raised (Fig. 1a). Consequently, it has proved difficult to develop adsorbents with the higher usable capacities needed for a commercially viable ANG storage system<sup>9</sup>. In pursuit of a new strategy for boosting usable capacity, we endeavoured to design an adsorbent with an ‘S-shaped’ or ‘stepped’ CH<sub>4</sub> adsorption isotherm, where the amount of CH<sub>4</sub> adsorbed would be small at low pressures but rise sharply just before the pressure reaches the desired storage pressure (Fig. 1b). Stepped isotherms have been observed for many flexible metal–organic frameworks that exhibit ‘gate-opening’ behaviour,

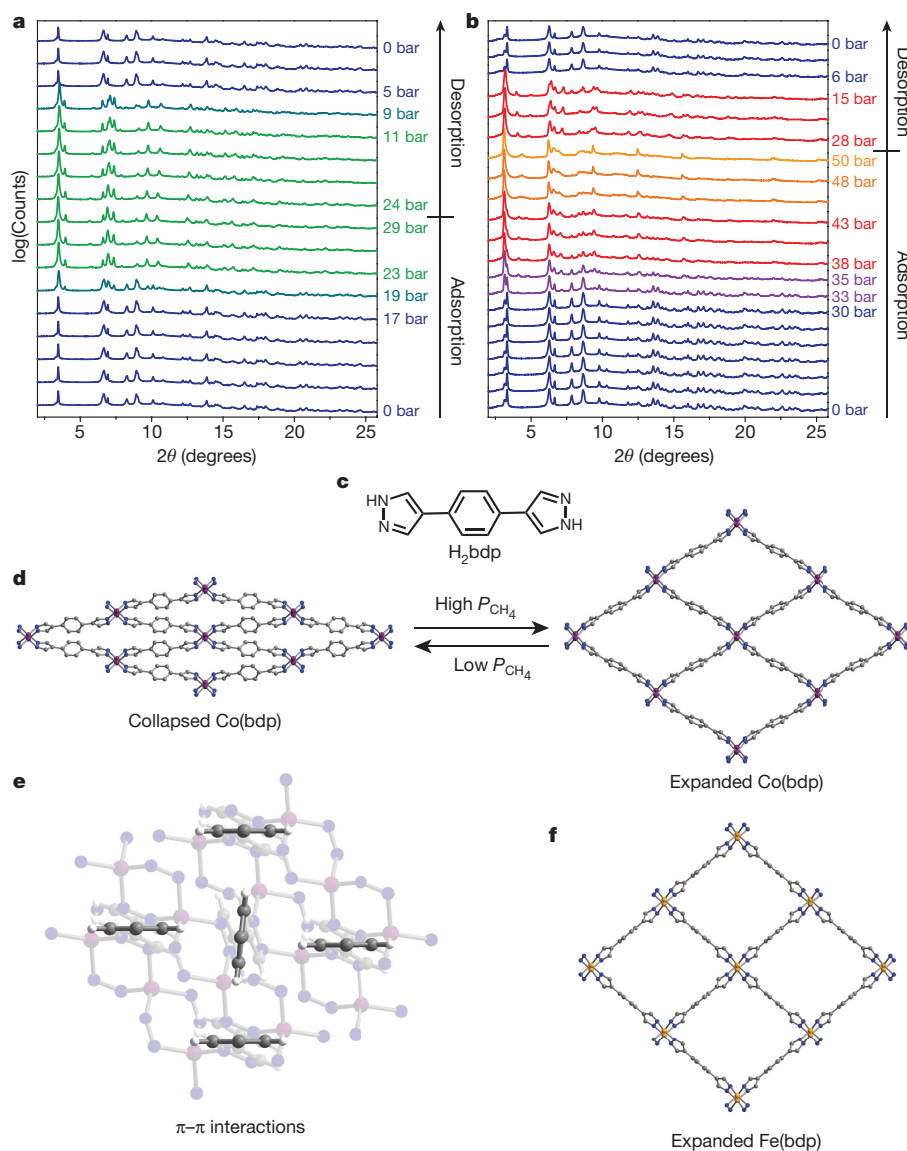
whereby a non-porous structure expands to a porous structure after a certain threshold gas pressure is reached, but none of these materials have exhibited characteristics beneficial for CH<sub>4</sub> storage applications<sup>12–16</sup>. If, however, a responsive adsorbent could be designed to expand to store a high density of CH<sub>4</sub> at 35–65 bar, and to collapse to push out all adsorbed CH<sub>4</sub> at a pressure near 5.8 bar, then it should be possible to reach higher usable capacities than have been realized for classical adsorbents.

The metal–organic framework Co(bdp) was selected as a potential responsive adsorbent for methane storage, owing to its large internal surface area and its previously demonstrated high degree of flexibility<sup>17</sup>. In its solvated form, this framework features one-dimensional chains of tetrahedral Co<sup>2+</sup> cations bridged by  $\mu^2$ -pyrazolates to form a structure with square channels with edge lengths of 13 Å. The N<sub>2</sub> adsorption isotherm of the evacuated framework at 77 K exhibits five distinct steps, which have been attributed to four structural transitions as the framework



**Figure 1 | High-pressure CH<sub>4</sub> adsorption isotherms.** **a, b**, The usable capacity is compared for an idealized adsorbent exhibiting a classical Langmuir-type adsorption isotherm (**a**) and an ‘S-shaped’ or ‘stepped’ adsorption isotherm (**b**), with the minimum desorption pressure  $P_{\text{des}}$  and the maximum adsorption pressure  $P_{\text{ads}}$  indicated by the vertical dashed grey lines. **c, d**, Total CH<sub>4</sub> adsorption isotherms for Co(bdp) (**c**) and Fe(bdp) (**d**) at 25 °C. Here  $P_{\text{des}} = 5.8$  bar and  $P_{\text{ads}} = 35$  bar are indicated by dashed grey lines. Filled circles represent adsorption; open circles represent desorption.

<sup>1</sup>Department of Chemistry, University of California, Berkeley, California 94720, USA. <sup>2</sup>Center for Neutron Research, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, USA. <sup>3</sup>Aix-Marseille University, CNRS Laboratoire MADIREL (UMR 7246), Centre de Saint Jérôme, 13397 Marseille Cedex 20, France. <sup>4</sup>Laboratory for Synchrotron Radiation – Condensed Matter, Swiss Light Source, Paul Scherrer Institute, CH-5232 Villigen, Switzerland. <sup>5</sup>Istituto di Cristallografia, Consiglio Nazionale delle Ricerche, and To.Sca.Lab., via Valleggio 11, 22100 Como, Italy. <sup>6</sup>Chemical and Biomolecular Engineering, University of Delaware, Newark, Delaware 19716, USA. <sup>7</sup>Dipartimento di Scienza e Alta Tecnologia, Università dell’Insubria, and To.Sca.Lab., via Valleggio 11, 22100 Como, Italy.



**Figure 2 | Powder X-ray diffraction and solid-state structures.**

**a, b,** Powder X-ray diffraction patterns ( $2\theta$  is the diffraction angle) are shown for Co(bdp) (**a**) and Fe(bdp) (**b**) at 25 °C and variable CH<sub>4</sub> pressures (as indicated), with X-ray wavelengths of 0.75009 Å and 0.72768 Å, respectively. For Co(bdp), the blue and green patterns correspond to the collapsed and expanded phases, respectively, with teal indicating patterns in which both phases are present during the transition between collapsed and expanded. For Fe(bdp), the blue and red patterns correspond to the collapsed and 40-bar expanded phases, respectively, with purple indicating patterns in which both phases are present during the transition from

collapsed to 40-bar expanded; orange patterns correspond to the 50-bar expanded phase. **c, d,** The bridging ligand precursor H<sub>2</sub>bdp (**c**) along with the crystal structures (**d**) of the collapsed (0 bar, 'low  $P_{\text{CH}_4}$ ') and CH<sub>4</sub>-expanded (30 bar, 'high  $P_{\text{CH}_4}$ ') phases of Co(bdp). **e,** Each benzene ring in the collapsed phase of Co(bdp) has four edge-to-face  $\pi$ - $\pi$  interactions with neighbouring benzene rings. **f,** Crystal structure of the CH<sub>4</sub>-expanded (40 bar) phase of Fe(bdp). In **d-f**, Purple, orange, grey, blue, and white spheres represent Co, Fe, C, N, and H atoms, respectively; some H atoms are omitted for clarity.

progresses from a collapsed phase with minimal porosity to a maximally expanded phase with a Langmuir surface area of 2,911 m<sup>2</sup> g<sup>-1</sup> (ref. 18).

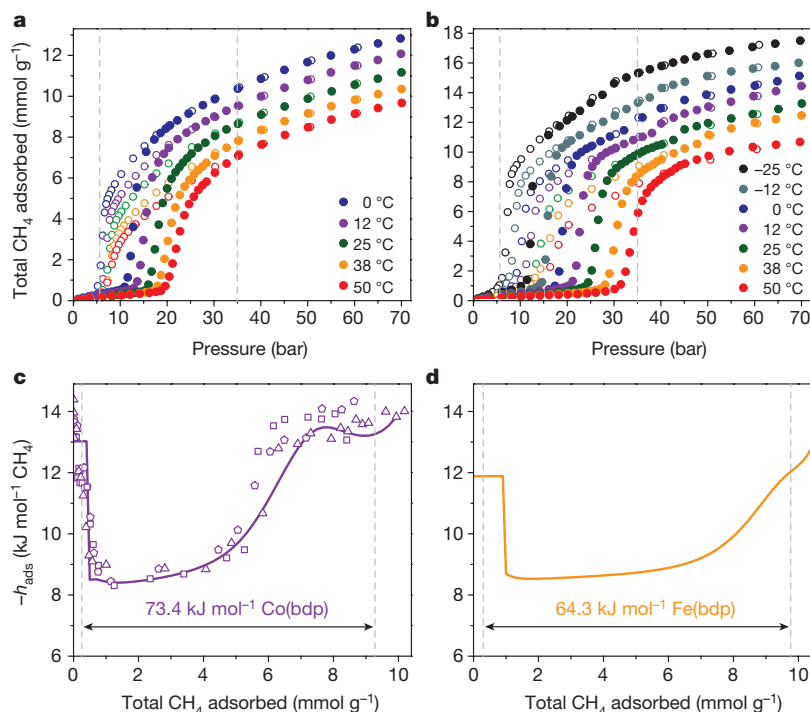
To investigate the ANG storage potential of Co(bdp), a high-pressure CH<sub>4</sub> adsorption isotherm was measured at 25 °C (Fig. 1c). There is minimal CH<sub>4</sub> uptake at low pressures and a sharp step in the adsorption isotherm at 16 bar. Although there is hysteresis in the desorption isotherm, the hysteresis loop is closed by 7 bar, such that there is less than 0.2 mmol g<sup>-1</sup> of CH<sub>4</sub> adsorbed at pressures below 5.8 bar. The step in the CH<sub>4</sub> isotherm is fully reproducible over at least 100 adsorption-desorption cycles (Extended Data Fig. 1), and can be attributed to a reversible structural phase transition between a collapsed, non-porous framework and an expanded, porous framework at transition pressures that are ideal for ANG storage.

To determine the specific structural changes responsible for the stepped CH<sub>4</sub> adsorption isotherm of Co(bdp), *in situ* powder X-ray

diffraction experiments were performed under various pressures of CH<sub>4</sub> at 25 °C. Under vacuum, only one crystalline phase is observed in the diffraction pattern, consistent with the complete conversion of Co(bdp) to a collapsed phase upon desolvation. From 17 bar to 23 bar, there are substantial changes to both the positions and intensities of the diffraction peaks, as peaks corresponding to the collapsed phase decrease in intensity and peaks corresponding to a new expanded phase increase in intensity (Fig. 2a). During desorption, this expanded phase is fully converted back to the collapsed phase between 10 bar and 5 bar.

Owing to the anisotropic peak widths and complex peak shapes that result from paracrystallinity effects<sup>19</sup>, analysis of the powder diffraction data is not trivial, but *ab initio* structure solutions followed by Rietveld refinements (Extended Data Fig. 7) were successfully performed using the diffraction data at 0 bar and 30 bar to provide crystal structures of the collapsed and expanded phases of Co(bdp) (Fig. 2d). As discussed





**Figure 3 | Variable-temperature equilibrium isotherms and differential enthalpies.** **a, b,** Total CH<sub>4</sub> adsorption isotherms at various temperatures for Co(bdp) (**a**) and Fe(bdp) (**b**), where a minimum desorption pressure of 5.8 bar and a maximum adsorption pressure of 35 bar are indicated by dashed grey lines. Filled circles represent adsorption; open circles represent desorption. **c,** Differential enthalpies of CH<sub>4</sub> adsorption ( $h_{\text{ads}}$ ) for Co(bdp),

as determined from variable-temperature adsorption isotherms (purple line) and three separate microcalorimetry experiments (open symbols). **d,** Differential enthalpies of CH<sub>4</sub> adsorption ( $h_{\text{ads}}$ ) for Fe(bdp), as determined from variable-temperature adsorption isotherms. Dashed grey lines in **c** and **d** indicate the amount of CH<sub>4</sub> adsorbed at 5.8 bar and 35 bar.

in the Supplementary Information and shown in Extended Data Fig. 8, paracrystallinity arises from highly correlated shifts of the positions of Co-pyrazolate chains in the crystallographic *a*–*b* plane, whereby neighbouring chains exhibit average displacements of approximately 0.5 Å from their average periodic positions. Importantly, this minor systematic disordering has no effect on the accuracy of the average crystal structures or the calculated crystallographic densities of each phase. Additionally, a substantial diffuse-scattering component is present in the experimental diffraction patterns, particularly at high CH<sub>4</sub> loadings. Although most of the diffuse scattering can be attributed to the thick-walled quartz glass capillaries used as sample holders in the diffraction experiments at high CH<sub>4</sub> pressures (Supplementary Fig. 11), there may also be some diffuse scattering that is intrinsic to Co(bdp), which could arise from minor local disorder or from scattering by adsorbed CH<sub>4</sub> molecules.

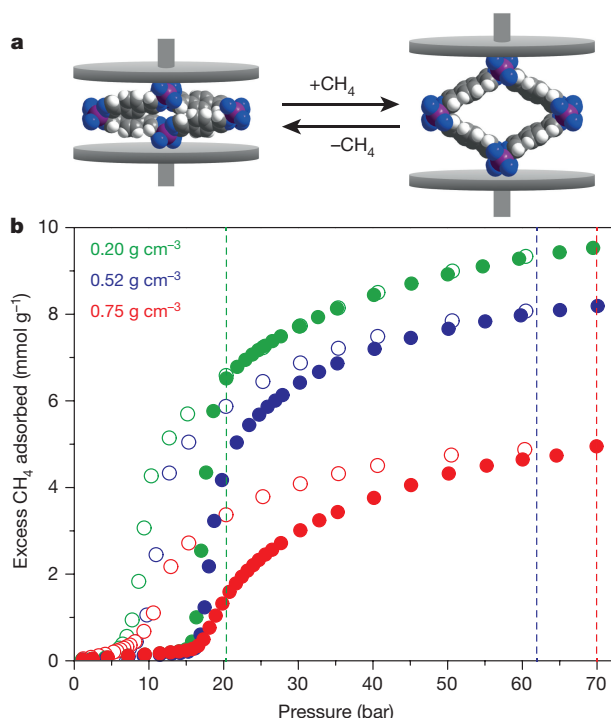
Even though the density of the collapsed phase (1.50 g cm<sup>-3</sup>) is nearly double that of the expanded phase (0.77 g cm<sup>-3</sup>), the Co<sup>2+</sup> ions adopt a similar pseudotetrahedral geometry in both structures. During the phase transition, the angles between the planes of the pyrazolate rings and the Co–N bonds decrease as the framework expands (Extended Data Fig. 6). In addition, the central benzene ring of the bdp<sup>2-</sup> ligand twists out of the plane of the two pyrazolates by 25° in the collapsed structure, resulting in edge-to-face  $\pi$ – $\pi$  interactions with four neighbouring benzene rings that probably provide most of the thermodynamic driving force for the collapse of Co(bdp) at low pressures (Fig. 2e)<sup>20</sup>. These close contacts between neighbouring bdp<sup>2-</sup> ligands lead to no accessible porosity, and thus no CH<sub>4</sub> adsorption, in the collapsed phase.

The usable CH<sub>4</sub> capacity of Co(bdp) at 25 °C is 155 cm<sup>3</sup> STP cm<sup>-3</sup> (v/v) for adsorption at 35 bar and 197 v/v for adsorption at 65 bar, which are the highest values of usable CH<sub>4</sub> capacity reported so far for any adsorbent under these conditions. A recent computational analysis of a database containing over 650,000 classical adsorbents predicted a theoretical-maximum 65-bar usable capacity of 196 v/v (ref. 9); however,

all adsorbents in this large-scale computational screening were rigid, and the potential utility of flexible adsorbents for CH<sub>4</sub> storage was not considered. The Co(bdp) usable capacities reported here are a result of the transition from the expanded to the collapsed phase leading to near complete CH<sub>4</sub> desorption by 5.8 bar. For comparison, the highest previously measured 35-bar and 65-bar usable capacities for any adsorbent are 143 v/v and 189 v/v, as obtained for the metal–organic frameworks HKUST-1 and UTSA-76a, respectively<sup>7,8,21</sup>. Both of these Cu<sub>2</sub> paddlewheel-based frameworks have high densities of CH<sub>4</sub> adsorption sites with a near-optimal (for maximizing the usable CH<sub>4</sub> capacity for ambient temperature adsorption at pressures between 35 bar and 65 bar) binding enthalpy of –15 kJ mol<sup>-1</sup> to –17 kJ mol<sup>-1</sup>, but display Langmuir-type adsorption isotherms that leave a substantial amount of unusable CH<sub>4</sub> adsorbed at 5.8 bar.

One major, and often overlooked, challenge in developing adsorbents for natural gas storage, or indeed for any gas storage application, involves managing the exothermic heat of adsorption and endothermic heat of desorption, both of which reduce the usable capacity of an adsorbent. These heat effects can be substantial, with temperature changes of as much as 80 °C observed during testing of prototype activated carbon-based ANG systems, and result in large reductions in the usable CH<sub>4</sub> capacity<sup>22,23</sup>. On-board thermal-management systems are essential to minimizing the negative impacts of the heats of sorption, but these engineering controls take up already limited space on a vehicle and add considerable cost and complexity<sup>24</sup>.

Responsive adsorbents, such as Co(bdp), offer the possibility of managing heat intrinsically within a material, rather than through an external system, by using the enthalpy change of a phase transition to partially, or perhaps even fully, offset the heats of sorption. For Co(bdp), the expansion of the framework during adsorption is endothermic, because energy is needed to overcome the greater thermodynamic stability of the collapsed phase. As a result, some of the enthalpy of CH<sub>4</sub> adsorption should go towards providing the heat needed for



**Figure 4 | Effect of mechanical pressure on CH<sub>4</sub> storage in Co(bdp).** **a**, Space-filling models of collapsed (left) and CH<sub>4</sub>-expanded (right) Co(bdp); purple, grey, blue, and white spheres represent Co, C, N, and H atoms, respectively. **b**, Excess CH<sub>4</sub> adsorption isotherms for Co(bdp) at 25 °C with different levels of applied external mechanical pressure, indicated by the inset, colour-coded bulk powder densities, with higher densities corresponding to greater applied mechanical pressure. The maximum CH<sub>4</sub> pressure for which hysteresis is still present is indicated for each bulk density by the appropriately coloured dashed line. Filled circles represent adsorption; open circles represent desorption.

the transition to the expanded phase, lowering the overall amount of heat released compared to adsorption in the absence of a phase transition. Similarly, the transition to the collapsed phase is exothermic, and some of the heat released by the framework as it collapses should offset the endothermic desorption of CH<sub>4</sub>.

In classical porous materials, low-coverage differential CH<sub>4</sub> adsorption enthalpies are generally  $-12 \text{ kJ mol}^{-1} \text{ CH}_4$  to  $-15 \text{ kJ mol}^{-1} \text{ CH}_4$  for adsorbents that do not have any strong CH<sub>4</sub> binding sites and are closer to  $-15 \text{ kJ mol}^{-1}$  to  $-25 \text{ kJ mol}^{-1}$  for adsorbents with the highest volumetric CH<sub>4</sub> capacities<sup>7,8</sup>. For the steepest region of the CH<sub>4</sub> adsorption isotherm of Co(bdp), the differential enthalpy is considerably lower, at just  $-8.4(3) \text{ kJ mol}^{-1}$  (where the uncertainty corresponds to  $\pm 1$  standard deviation), because the endothermic framework expansion partially offsets the exothermic heat of adsorption (Fig. 3c). After the transition to the expanded Co(bdp) phase is complete, the differential enthalpy approaches  $-13 \text{ kJ mol}^{-1}$ , which is consistent with weak CH<sub>4</sub> physical adsorption in the absence of a phase transition to mitigate heat. To confirm the accuracy of the calculated differential enthalpies, the heat released during CH<sub>4</sub> adsorption was directly measured by performing variable-pressure microcalorimetry experiments. As shown in Fig. 3c, the differential enthalpies obtained from calorimetry are in excellent agreement with those calculated from the variable-temperature adsorption isotherms.

The total amount of heat released when increasing the pressure of CH<sub>4</sub> adsorbed in Co(bdp) from 5.8 bar to 35 bar, as would occur during refuelling of an ANG vehicle, is calculated by integrating the differential-enthalpy curve with respect to the amount of CH<sub>4</sub> adsorbed. The  $73.4 \text{ kJ}$  of heat released per litre of Co(bdp) represents a 33% reduction relative to the  $109 \text{ kJ l}^{-1}$  of heat released by HKUST-1 under the same conditions, even though the amount of CH<sub>4</sub> adsorbed in Co(bdp)

is 8% greater. We further calculate that  $93.9 \text{ kJ l}^{-1}$  of heat would be released for hypothetical CH<sub>4</sub> adsorption in a rigid Co(bdp) framework—28% higher than when adsorption occurs with a phase transition to provide heat mitigation<sup>25</sup>.

By chemically modifying Co(bdp), we hypothesized that it might be possible to obtain a new flexible framework with a similar stepped CH<sub>4</sub> isotherm, but a higher-energy phase transition that could provide even greater intrinsic heat management. Because one-dimensional chains are known to form with tetrahedral Fe<sup>2+</sup> ions bridged by  $\mu^2$ -pyrazolates<sup>26</sup>, we anticipated that it might be possible to synthesize an isostructural iron analogue of Co(bdp). By heating FeCl<sub>2</sub> and H<sub>2</sub>bdp in a mixture of *N,N*-dimethylformamide (DMF) and methanol, we indeed obtained Fe(bdp) as yellow, block-shaped crystals. X-ray analysis of a DMF-solvated crystal (Extended Data Fig. 6) confirmed that Fe(bdp) is isostructural to Co(bdp). Fe(bdp) has a stepped high-pressure CH<sub>4</sub> isotherm at 25 °C (Fig. 1d), suggesting that this new compound also undergoes a reversible phase transition between a collapsed and expanded framework. Although the total CH<sub>4</sub> uptake is comparable to that of Co(bdp), the adsorption and desorption steps occur at the considerably higher pressures of 24 bar and 10 bar, respectively, suggesting that replacing Co with Fe increases the energy of the phase transition.

*In situ* powder X-ray diffraction experiments from 0 bar to 50 bar of CH<sub>4</sub> (Fig. 2b) and subsequent Rietveld refinements afforded the collapsed and CH<sub>4</sub>-expanded crystal structures of Fe(bdp). Although the collapsed phase is nearly identical to that of Co(bdp), with edge-to-face  $\pi$ - $\pi$  interactions and no accessible porosity, the volume of the expanded Fe(bdp) phase at 40 bar is 9% greater than that of Co(bdp) (Fig. 2f). In contrast to Co(bdp), we observe a second transition for Fe(bdp) at pressures above 40 bar, wherein Fe(bdp) slightly expands to a framework with nearly perfect square channels (Extended Data Fig. 6). In spite of its greater expansion and lower crystallographic density, the usable CH<sub>4</sub> capacity of Fe(bdp) is still higher than all known adsorbents at 150 v/v and 190 v/v for 35 bar and 65 bar adsorption, respectively.

Although Fe(bdp) and Co(bdp) have similar usable capacities, the initial Fe(bdp) phase transition offsets more heat, and only  $64.3 \text{ kJ}$  of heat is released per litre of adsorbent during CH<sub>4</sub> adsorption at 35 bar, which is 12% lower than for Co(bdp) and 41% lower than for HKUST-1. This is a direct consequence of the larger increase in the enthalpy of Fe(bdp) ( $8.1 \text{ kJ mol}^{-1}$ ) than of Co(bdp) ( $7.0 \text{ kJ mol}^{-1}$ ) during the phase transition, which mitigates more heat of adsorption, thereby providing a greater source of intrinsic thermal management. This result demonstrates how a slight variation in the metal-organic framework can be used to improve its intrinsic thermal management, and it is very likely that similar effects will prove possible through alteration of the bdp<sup>2-</sup> bridging ligand.

Examining the temperature dependence of the CH<sub>4</sub> isotherms of Co(bdp) and Fe(bdp) (Extended Data Figs 2, 3) reveals another advantage of these materials, involving a reduction in the effect of cooling during desorption. Consistent with other gate-opening metal-organic frameworks, the CH<sub>4</sub> adsorption and desorption steps in Co(bdp) and Fe(bdp) shift to lower pressures at lower temperatures (Fig. 3a, b). As long as the temperature stays above 0 °C in Co(bdp) or  $-25 \text{ °C}$  in Fe(bdp), however, the transition to the collapsed phase occurs above 5.8 bar, and the usable CH<sub>4</sub> capacity will not be affected by cooling (Supplementary Tables 2, 3). This property has practical benefits for driving in cold-weather climates and should further reduce the overall thermal management required in an ANG system.

Recent work<sup>27–29</sup> has shown that it is possible to induce a phase transition in flexible metal-organic frameworks by applying external mechanical pressure. With this in mind, we proposed that applying moderate mechanical pressure could provide a means of further tuning the CH<sub>4</sub> adsorption and desorption step pressures in Co(bdp) and Fe(bdp) and of increasing the energy of the phase transition to offset more heat. To investigate this concept, high-pressure CH<sub>4</sub> adsorption isotherms were measured for Co(bdp) at different levels of applied uniaxial mechanical pressure.

At higher mechanical pressures and higher compaction densities, both the adsorption and desorption isotherm steps shift to higher CH<sub>4</sub> pressures, which is consistent with an increase in the energy of the phase transition (Fig. 4). In addition, the isotherm hysteresis loop remains open until higher CH<sub>4</sub> pressures, with hysteresis observed to pressures of at least 70 bar for the highest applied mechanical pressure. Because hysteresis at a given pressure implies that a phase transition is still occurring<sup>30</sup>, this result suggests that some Co(bdp) crystallites are expanding at much higher CH<sub>4</sub> pressures when under an applied external mechanical pressure. Because Co(bdp) crystallites in a bulk powder will be at different orientations with respect to the direction of uniaxial compression (Extended Data Fig. 4), there will be a distribution of local mechanical pressures experienced by different crystallites. Crystallites that experience higher external pressures will have a greater free energy change associated with the phase transition and will open at higher pressures<sup>31</sup>. Overall, these results present the prospect of using mechanical work, such as provided through an elastic bladder, as a means of thermal management in a gas-storage system based on a flexible adsorbent.

Designing new flexible adsorbents with stronger gas binding sites and higher-energy phase transitions provides a promising route to achieving even higher usable capacities and greater intrinsic heat management in a next generation of gas-storage materials. Moreover, improved compaction and packing strategies should allow further reductions to external thermal-management requirements and optimization of the overall storage-system performance.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 28 April; accepted 4 September 2015.**

**Published online 26 October 2015.**

- Service, R. F. Stepping on the gas. *Science* **346**, 538–541 (2014).
- Yeh, S. An empirical analysis on the adoption of alternative fuel vehicles: the case of natural gas vehicles. *Energy Policy* **35**, 5865–5875 (2007).
- Whyatt, G. A. *Issues Affecting Adoption of Natural Gas Fuel in Light- and Heavy-Duty Vehicles*. Report No. PNNL-19745 (US Department of Energy, 2010).
- Wegrzyn, J. & Gurevich, M. Adsorbent storage of natural gas. *Appl. Energy* **55**, 71–83 (1996).
- Makal, T. A., Li, J.-R., Lu, W. & Zhou, H.-C. Methane storage in advanced porous materials. *Chem. Soc. Rev.* **41**, 7761–7779 (2012).
- He, Y., Zhou, W., Qian, G. & Chen, B. Methane storage in metal-organic frameworks. *Chem. Soc. Rev.* **43**, 5657–5678 (2014).
- Peng, Y. *et al.* Methane storage in metal-organic frameworks: current records, surprise findings, and challenges. *J. Am. Chem. Soc.* **135**, 11887–11894 (2013).
- Mason, J. A., Veenstra, M. & Long, J. R. Evaluating metal-organic frameworks for natural gas storage. *Chem. Sci.* **5**, 32–51 (2014).
- Simon, C. M. *et al.* The materials genome in action: identifying the performance limits for methane storage. *Energy Environ. Sci.* **8**, 1190–1199 (2015).
- Advanced Research Projects Agency – Energy. *Methane Opportunities for Vehicular Energy* (Funding opportunity no. DE-FOA-0000672, US Department of Energy, 2012).
- Noguchi, H. *et al.* Clathrate-formation mediated adsorption of methane on Cu-complex crystals. *J. Phys. Chem. B* **109**, 13851–13853 (2005).
- Horike, S., Shimomura, S. & Kitagawa, S. Soft porous crystals. *Nature Chem.* **1**, 695–704 (2009).
- Férey, G. & Serre, C. Large breathing effects in three-dimensional porous hybrid matter: facts, analyses, rules and consequences. *Chem. Soc. Rev.* **38**, 1380–1399 (2009).
- Schneemann, A. *et al.* Flexible metal-organic frameworks. *Chem. Soc. Rev.* **43**, 6062–6096 (2014).
- Li, D. & Kaneko, K. Hydrogen bond-regulated microporous nature of copper complex-assembled microcrystals. *Chem. Phys. Lett.* **335**, 50–56 (2001).
- Kitaura, R., Seki, K., Akiyama, G. & Kitagawa, S. Porous coordination-polymer crystals with gated channels specific for supercritical gases. *Angew. Chem. Int. Edn* **42**, 428–431 (2003).

- Choi, H. J., Dincă, M. & Long, J. R. Broadly hysteretic H<sub>2</sub> adsorption in the microporous metal-organic framework Co(1,4-benzenedipyrzolate). *J. Am. Chem. Soc.* **130**, 7848–7850 (2008).
- Salles, F. *et al.* Multistep N<sub>2</sub> breathing in the metal-organic framework Co(1,4-benzenedipyrzolate). *J. Am. Chem. Soc.* **132**, 13782–13788 (2010).
- Hosemann, R. & Bagchi, S. N. *Direct Analysis of Diffraction by Matter* (North-Holland, 1962).
- Sinnokrot, M. O., Valeev, E. F. & Sherrill, C. D. Estimates of the *ab initio* limits for  $\pi$ - $\pi$  interactions: the benzene dimer. *J. Am. Chem. Soc.* **124**, 10887–10893 (2002).
- Li, B. *et al.* A porous metal-organic framework with dynamic pyrimidine groups exhibiting record high methane storage working capacity. *J. Am. Chem. Soc.* **136**, 6207–6210 (2014).
- Barbosa Mota, J. P., Rodrigues, A. E., Saatdjian, E. & Tondeur, D. Dynamics of natural gas adsorption storage systems employing activated carbon. *Carbon* **35**, 1259–1270 (1997).
- Walton, K. S. & LeVan, M. D. Natural gas storage cycles: influence of nonisothermal effects and heavy alkanes. *Adsorption* **12**, 227–235 (2006).
- Weickert, M., Marx, S., Müller, U. & Arnold, L. Sorption store for gas with multiple adsorbent media. World Intellectual Property Organization patent WO 2015/022633 A1 (2015).
- Coudert, F.-X., Jeffroy, M., Fuchs, A. H., Boutin, A. & Mellot-Draznieks, C. Thermodynamics of guest-induced structural transitions in hybrid organic-inorganic frameworks. *J. Am. Chem. Soc.* **130**, 14294–14302 (2008).
- Patrick, B. O., Reif, W. M., Sánchez, V., Storr, A. & Thompson, R. C. Polybis(pyrazolato)iron(II) and poly-2,2'-bipyridinetetrakis(imidazolato)-diiron(II) and -dicobalt(II): from short-range magnetic interactions in the pyrazolate to long-range ferromagnetic ordering in the imidazoles. *Polyhedron* **20**, 1577–1585 (2001).
- Beurroies, I. *et al.* Using pressure to provoke the structural transition of metal-organic frameworks. *Angew. Chem. Int. Edn* **49**, 7526–7529 (2010).
- Yot, P. G. *et al.* Large breathing of the MOF MIL-47(V<sup>IV</sup>) under mechanical pressure: a joint experimental-modelling exploration. *Chem. Sci.* **3**, 1100–1104 (2012).
- Coudert, F.-X. Responsive metal-organic framework materials: under pressure, taking the heat, in the spotlight, with friends. *Chem. Mater.* **27**, 1905–1916 (2015).
- Neimark, A. V., Coudert, F.-X., Boutin, A. & Fuchs, A. H. Stress-based model for the breathing of metal-organic frameworks. *J. Phys. Chem. Lett.* **1**, 445–449 (2010).
- Ghysels, A. *et al.* On the thermodynamics of framework breathing: a free energy model for gas adsorption in MIL-53. *J. Phys. Chem. C* **117**, 11540–11554 (2013).

**Supplementary information** is available in the online version of the paper.

**Acknowledgements** This research was supported by the Advanced Research Projects Agency – Energy (ARPA-E) of the US Department of Energy (DoE). Powder X-ray diffraction data were collected at beamline 17-BM-B at the Advanced Photon Source, a DoE Office of Science User Facility operated by Argonne National Laboratory under contract no. DE-AC02-06CH11357 and at beamline MS-X04SA of the Swiss Light Source (SLS) at the Paul Scherrer Institut. Single-crystal X-ray diffraction experiments were performed at beamline 11.3.1 at the Advanced Light Source, a DoE Office of Science User Facility operated by Lawrence Berkeley National Laboratory under contract no. DE-AC02-05CH11231. In addition, we thank M. Veenstra, D. A. Boysen, T. M. McDonald, D. J. Xiao, M. Nippe, Z. Hulvey, G. J. Halper, K. J. Gagnon, S. J. Teat, and the technical staff of the MS-X04SA beamline at SLS for experimental assistance and discussions. We also thank the National Science Foundation for providing graduate fellowship support for J.O. and J.E.B.

**Author Contributions** J.A.M. and J.R.L. formulated the project. J.A.M., J.O., and M.K.T. synthesized the compounds and collected the gas adsorption data. J.A.M. analysed all adsorption data. J.A.M., J.O., M.R.H., C.M.B., A.C., A.G., and N.M. collected and analysed the powder X-ray diffraction data. J.O. and M.I.G. collected and analysed the single-crystal X-ray diffraction data. J.E.B. collected all SEM images. J.A.M. and J.E.B. performed the thermodynamics calculations, and J.R. and P.L.L. performed the microcalorimetry measurements. J.A.M. performed all mechanical pressure experiments, with assistance from J.E.B. J.A.M. and J.R.L. wrote the paper, and all authors contributed to revising the paper.

**Author Information** Metrical data for the solid-state structures of collapsed Co(bdp), expanded Co(bdp), collapsed Fe(bdp), 40-bar expanded Fe(bdp), 50-bar expanded Fe(bdp), DMF-solvated Fe(bdp) at 100 K, and DMF-solvated Fe(bdp) at 300 K are available free of charge from the Cambridge Crystallographic Data Centre under reference numbers CCDC 1058444–1058450. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.R.L. ([jrlong@berkeley.edu](mailto:jrlong@berkeley.edu)).



## METHODS

**Materials.** Anhydrous *N,N*-dimethylformamide (DMF) was obtained from a JC Meyer solvent system. The ligand 1,4-benzenedipyrzole ( $H_2bdp$ ) was synthesized according to a literature procedure<sup>17</sup>. All other reagents were obtained from commercial vendors and used without further purification. Ultra-high purity grade (99.999% purity) helium, dinitrogen, and methane were used for all adsorption measurements.

**Synthesis of Co(bdp).** The compound Co(bdp) was synthesized using a strategy adopted from a previous report<sup>17</sup>. Specifically, a 500-ml solvent bomb was charged with a magnetic stirring bar,  $Co(CF_3SO_3)_2$  (4.96 g, 0.0139 mol),  $H_2bdp$  (2.46 g, 0.0117 mol), and *N,N*-diethylformamide (90 ml). The reaction mixture was degassed by the freeze–pump–thaw method for 5 cycles then sealed by closing the stopcock of the solvent bomb while the frozen reaction mixture was still under vacuum. The solvent bomb was then heated at 160 °C for 4.5 days to afford a purple microcrystalline solid. The solvent bomb was backfilled with  $N_2$ , and the solid was collected by filtration. Before drying, the wet solid powder was immediately transferred to a 500-ml glass jar, and 400 ml of DMF was added. The jar was heated at 120 °C for 12 h, then cooled to room temperature. The DMF was decanted and replaced with 400 ml of fresh DMF. The jar was reheated at 120 °C, followed by decanting and replacing with fresh DMF. This was repeated four additional times. The DMF was then decanted and replaced with dichloromethane (DCM). The DCM was partially decanted until 50 ml of solution was remaining. The resultant slurry was transferred to a 100-ml Schlenk flask, and the DCM was evaporated by flowing  $N_2$  at room temperature. The resultant solid was dried by flowing  $N_2$  at 160 °C for 12 h, then placed under dynamic vacuum at 160 °C for 24 h. The activated solid was immediately transferred to a glovebox and handled under a  $N_2$  atmosphere for all further experiments.

**Synthesis of Fe(bdp).** In a glovebox under a  $N_2$  atmosphere,  $H_2bdp$  (0.200 g, 0.95 mmol) in DMF (9 ml) was heated to 120 °C while stirring for 20 min in a 20 ml glass vial. The resultant yellow suspension was cooled. A solution of  $FeCl_2$  (0.197 g, 1.55 mmol) in methanol (1 ml) was added to the cooled suspension of  $H_2(bdp)$  in DMF, and the vial was sealed and heated at 120 °C while stirring. The hot, orange–yellow solution yielded a yellow microcrystalline powder after several hours. Samples suitable for gas adsorption studies were prepared using multiple vials of the same reaction scale in a glovebox under a  $N_2$  atmosphere and by washing the resultant material nine times with hot DMF ( $9 \times 18$  ml), before drying under high vacuum at 170 °C for 24 h. The activated sample was handled under a  $N_2$  atmosphere for all further experiments. IR (neat,  $cm^{-1}$ ): 1,573 (s), 1,336 (w), 1,239 (s), 1,110 (s), 1,041 (s), 952 (s), 859 (s), 849 (s), 832 (s), 824 (s), 644 (s), 534 (s). Anal. Calcd for  $FeC_{12}H_8N_4$ : C, 54.58; H, 3.05; N, 21.22. Found: C, 54.18; H, 2.36; N, 20.67. To obtain single crystals suitable for X-ray diffraction, a 9:1 mixture of DMF and methanol was used to create solutions of  $FeCl_2$  (9.0 mg, 0.07 mmol in 0.1 ml solvent) and  $H_2bdp$  (4.0 mg, 0.019 mmol in 0.9 ml solvent). The  $FeCl_2$  solution and the  $H_2bdp$  solution were added together in a 4-ml vial. The vial was then sealed, and the clear yellow solution was heated at 120 °C for 24 h. Block-shaped yellow crystals formed on the sides of the vial after several hours.

**Low-pressure gas adsorption measurements.** Gas adsorption isotherms for pressures in the range of 0–1.1 bar were measured using a Micromeritics ASAP 2020 or 2420 instrument. Activated samples were transferred under a  $N_2$  atmosphere to preweighed analysis tubes, which were capped with a Transeal. Each sample was evacuated on the ASAP until the outgas rate was less than  $3 \mu\text{bar min}^{-1}$ . The evacuated analysis tube containing degassed sample was then carefully transferred to an electronic balance and weighed to determine the mass of sample (typically 100–200 mg). The tube was then fitted with an isothermal jacket and transferred back to the analysis port of the ASAP. The outgas rate was again confirmed to be less than  $3 \mu\text{bar min}^{-1}$ . Langmuir surface areas were determined by measuring  $N_2$  adsorption isotherms in a 77-K liquid  $N_2$  bath and calculated using the Micromeritics software, assuming a value of  $16.2 \text{ \AA}^2$  for the molecular cross-sectional area of  $N_2$ . The Langmuir surface areas of Co(bdp) and Fe(bdp) are  $2,911 \text{ m}^2 \text{ g}^{-1}$  and  $2,780 \text{ m}^2 \text{ g}^{-1}$ , respectively. Full 77-K  $N_2$  adsorption isotherms for Co(bdp) and Fe(bdp) can be found in Supplementary Fig. 1. Note that BET surface areas cannot be accurately determined for either framework because of the steps in the low-pressure region of the 77-K  $N_2$  adsorption isotherms.

**High-pressure  $CH_4$  adsorption measurements.** High-pressure  $CH_4$  adsorption isotherms in the range of 0–70 bar were measured on an HPVA-II-100 from Particulate Systems, a Micromeritics company. In a typical measurement, 0.5–1.0 g of activated sample was loaded into a tared stainless steel sample holder inside a glovebox under a  $N_2$  atmosphere. Prior to connecting the sample holder to the VCR fittings of the complete high-pressure assembly inside the glovebox, the sample holder was weighed to determine the sample mass. The sample holder was then transferred to the HPVA-II-100, connected to the instrument's analysis port via an OCR fitting, and evacuated at room temperature for at least 2 h. The

sample holder was placed inside an aluminium recirculating Dewar connected to a Julabo FP89-HL isothermal bath filled with Julabo Thermal C2 fluid. The temperature stability of the isothermal bath is  $\pm 0.02$  °C. Methods for accurately measuring the relevant sample freespaces, which involve the expansion of He from a calibrated volume at 0.7 bar and 25 °C to the evacuated sample holder, were described in detail previously<sup>8</sup>. Non-ideality corrections were performed using the  $CH_4$  compressibility factors tabulated in the NIST REFPROP database<sup>32</sup> at each measured temperature and pressure.

A sample size of 1.032 g was used for the 25-°C usable capacity calculations, compaction studies, and cycling studies with Co(bdp), whereas a sample size of 0.584 g was used for the variable-temperature measurements. For Fe(bdp), a sample size of 0.274 g was used for high-pressure adsorption measurements, with the exception of the isotherms measured at  $-12$  °C and  $-25$  °C for which a sample size of 0.322 g was used.

To determine the usable  $CH_4$  capacity of Co(bdp) and Fe(bdp), experimentally measured excess gravimetric adsorption data (Extended Data Fig. 3) were converted to total volumetric adsorption data using the pore volume and crystallographic density of the  $CH_4$ -expanded phases (see Supplementary Text for details). All usable capacity calculations assume a minimum desorption pressure of 5.8 bar. Although the minimum desorption pressure required for natural gas to flow from the adsorbent to the combustion engine can vary from 3.5 bar to 10 bar depending on the specific requirements of fuel injectors, filters, and other engine components, a value of 5.8 bar has been adopted by many groups<sup>9,10</sup> for initial materials comparisons.

**High-pressure  $CH_4$  adsorption measurements under applied mechanical pressure.** For the high-pressure  $CH_4$  adsorption measurements of Co(bdp) at different applied mechanical pressures, an aluminium sample holder was designed and used (Extended Data Fig. 5). The sample is loaded in the volume between the fritted and blank gaskets. The free volume between the fritted and blank gaskets in the absence of a sample was determined by expansion of He from a calibrated volume to be 5.242 ml. Initially, 1.032 g of Co(bdp) was loaded into this volume, resulting in a bulk density of  $0.197 \text{ g ml}^{-1}$  for the uncompacted powder. After measuring a high-pressure  $CH_4$  adsorption isotherm, the sample holder was returned to a glovebox under a  $N_2$  atmosphere, and the cell was opened by removing the cap behind the blank gasket. An aluminium rod with an outer diameter slightly less than the inner diameter of the sample holder was then inserted. A mechanical press was used to compact the sample by pushing down on the rod. A fresh blank gasket was then sealed behind the rod so that the rod was left pressed against the sample, with a continuously applied uniaxial mechanical pressure. The sample holder was returned to the high-pressure instrument and fully evacuated before measuring a high-pressure  $CH_4$  adsorption isotherm. This experiment was repeated after inserting additional metal rods to further compact the Co(bdp), increase the applied mechanical pressure, and reduce the sample volume. Packing densities for each experiment were calculated by subtracting the volume of each rod from the original sample volume.

The decrease in the total amount of  $CH_4$  adsorbed at higher mechanical pressures (Fig. 4) is not due to framework degradation, as is often observed when compacting classical adsorbents<sup>7</sup>, and can instead be explained by insufficient  $CH_4$  pressure to induce a phase transition in some crystallites and by a lack of sufficient free volume for all crystallites to expand into. To confirm this, a  $CH_4$  adsorption isotherm was measured after compacting collapsed Co(bdp) to a packing density of  $0.75 \text{ g cm}^{-3}$ , which is just below the crystallographic density of the expanded phase, and releasing the applied mechanical pressure by removing the metal rod. The resulting isotherm was found to be nearly identical to the pre-compaction isotherm, demonstrating that all Co(bdp) crystallites could once again fully expand (Extended Data Fig. 5).

**Powder X-ray diffraction measurements.** Powder X-ray diffraction data for Co(bdp) and Fe(bdp) were collected on beamline 17-BM-B at the Advanced Photon Source (APS) at Argonne National Laboratory and beamline MS-X04SA at the Swiss Light Source (SLS) at the Paul Scherrer Institut (Extended Data Fig. 7). For variable  $CH_4$  pressure experiments, approximately 10 mg of fully desolvated framework was loaded into 1.5-mm quartz glass capillaries inside a glovebox under a  $N_2$  atmosphere. Each capillary was attached to a custom-designed gas-dosing cell, which is equipped with a gas valve, and was then transferred to the goniometer head. All adsorbed  $N_2$  was removed by evacuating *in situ* using a turbomolecular pump. A cryostat was used to hold the temperature constant at 25 °C, and variable pressures of  $CH_4$  were dosed to the samples. Diffraction data were collected after allowing each dose to equilibrate for several minutes. All X-ray wavelengths were between 0.72 Å and 0.78 Å, and are specified for each experiment in the relevant figures and tables.

The structure solution and refinement procedure used in this study followed the standard protocol developed by us, and others, for the structural characterization

of polycrystalline samples of non-ideal crystallinity and moderately complex structures by *ab initio* powder diffraction methods. The specific details of the crystal structure determinations are discussed in the Supplementary Information, but the general procedure, which was fully adopted for the collapsed Co(bdp) phase, is summarized here. A standard peak search, followed by peak profile fitting was first used to determine accurate peak positions of several well-separated low-angle peaks. These peak positions were used to obtain approximate lattice parameters via the single-value-decomposition indexing procedure implemented in the software TOPAS-R (Bruker AXS, version 3.0, 2005), which were later refined by the structureless Le Bail method as implemented in TOPAS-R. Systematic absences, density considerations and previous knowledge of isotopic species coherently allowed the derivation of the correct space group, which was later confirmed by successful structure solution and Rietveld refinement. The structural model was derived using the simulated annealing procedure as implemented in TOPAS-R, which is a real-space structure-solution technique, with a single freely floating metal ion and an idealized half bdp<sup>2-</sup> ligand defined using *z*-matrix formalism. In the collapsed phase of Co(bdp), for instance, the metal atom was located on a two-fold axis at (0, *y*, 1/4), whereas the half bdp<sup>2-</sup> ligand was hinged about the inversion centre at (1/4, 3/4, 0). Once an initial structural model was established, complete Rietveld refinements were performed in the software TOPAS-R. The background was modelled with Chebyshev polynomials, and Lorentz and absorption correction factors were applied. A single isotropic *B* value was attributed to all atoms, and found to act, as expected, as a scavenger for  $\theta$ -dependent systematic errors, which are not suitably taken into account in the data-reduction process. After the retrieval of the lattice metrics and space-group symmetry for the CH<sub>4</sub>-expanded phases, defining a starting structural model was straightforward, because it is implicit in the isotopic nature of the compounds. The contribution of the (probably tumbling, but not necessarily randomly located) CH<sub>4</sub> molecules to the overall scattering power was neglected, which probably contributes to the decreased physical meaning of the atomic displacement parameter values, as is common for crystal structures determined from powder diffraction data.

As indicated in the main text, the peak widths of the collapsed Co(bdp) and Fe(bdp) could not be modelled by convoluting conventional Lorentzian and Gaussian functions (or their combinations) with systematic  $1/\cos(\theta)$  or  $\tan(\theta)$  dependency, respectively, or with smooth *hkl*-dependent models (such as spherical harmonics). Instead, we began by separately modelling the *hkl* peak widths as distinct from the axial reflections of the *h*00- and 0*k*0-type peak widths using a purely phenomenological model. We also developed a paracrystalline model for collapsed Co(bdp) (discussed in Supplementary Text) that is also representative of the collapsed Fe(bdp) phase.

**Single-crystal X-ray diffraction measurements.** X-ray diffraction analyses were performed on a single crystal of Fe(bdp) that was coated with Paratone-N oil and mounted on a MiTeGen loop. The crystal of Fe(bdp) was first kept frozen at 100 K by an Oxford Cryosystems Cryostream 800 plus, and after a full data collection, the crystal was warmed to 298 K for a second data collection. Diffraction data for Fe(bdp) was collected at beamline 11.3.1 at the Advanced Light Source, Lawrence Berkeley National Laboratory using synchrotron radiation (wavelength  $\lambda = 0.7749$  Å) with 1° omega scans for the 100-K structure, and 4° phi and 1° omega scans for the 298 K structure. A Bruker PHOTON100 CMOS diffractometer was used for data collection, and the corresponding Bruker AXS APEX II software was used for data collection and reduction. Raw data were integrated and corrected for Lorentz and polarization effects using the Bruker AXS SAINT software. Absorption corrections were applied using TWINABS for the 100-K structure and SADABS for the 298-K structure. Space-group assignments were

determined by examination of systematic absences, *E* statistics, and successive refinement of the structures of Fe(bdp) at 100 K and 298 K. The structures were solved using direct methods with SHELXS and refined using SHELXL operated in the OLEX2 interface. Thermal parameters were refined anisotropically for all non-hydrogen atoms. Hydrogen atoms were placed in ideal positions and refined using a riding model for all structures.

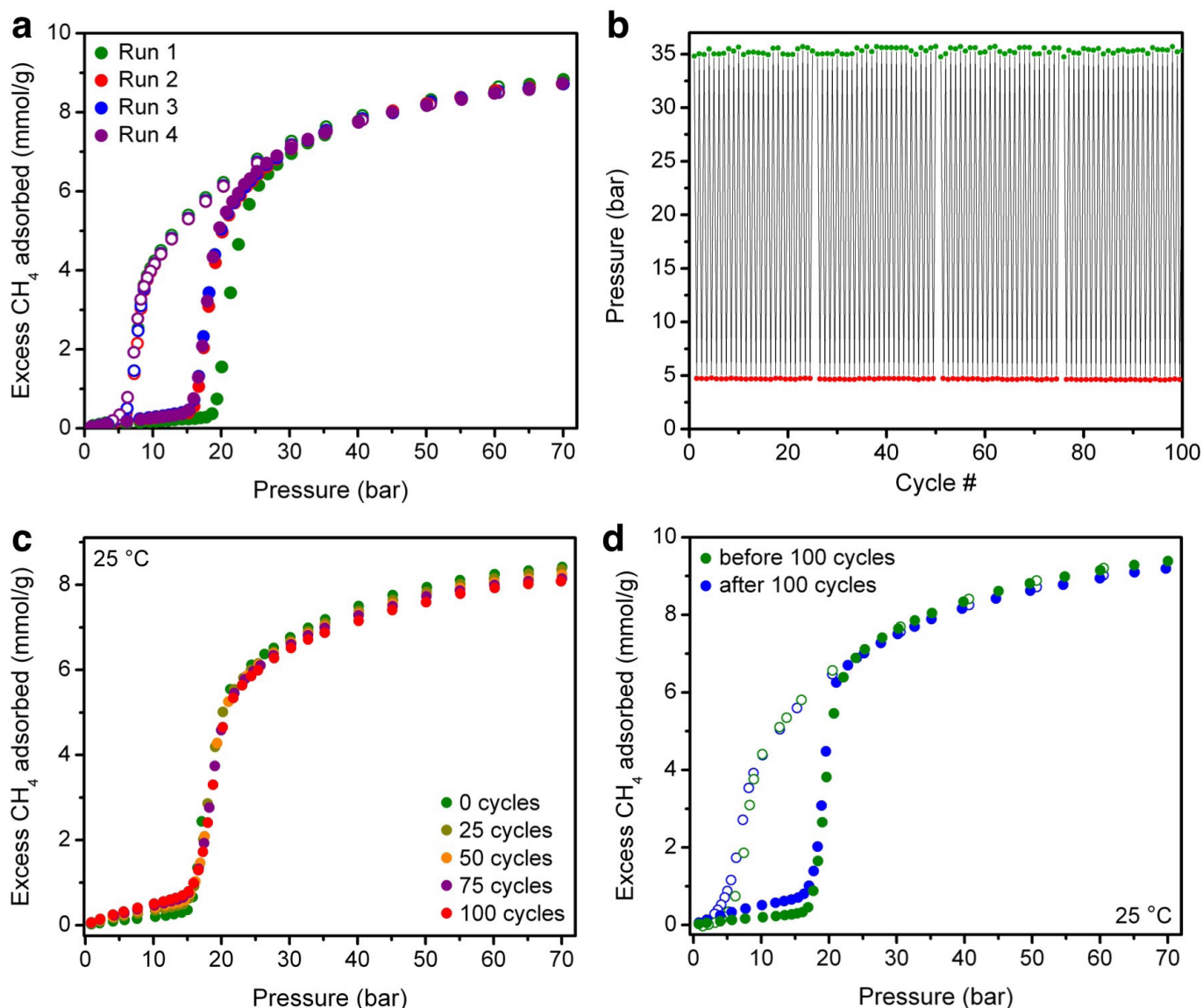
The crystal was determined to be twinned at 100 K and a suitable unit cell was determined that is similar to that previously reported for Co(bdp)•3DMF in the space group *P*<sub>2</sub><sub>1</sub>/*c* (ref. 33). The program CELL\_NOW was used to determine the orientation matrices, and the domains were found to be related by a 179.9° rotation around the reciprocal axis [0.5, 0, 1]. Raw data for both matrices were integrated and corrected for absorption using TWINABS. Solution and refinement of the data in *P*<sub>2</sub><sub>1</sub>/*c* required substantially fewer restraints in structure refinement and gave much lower values for *R*<sub>1</sub> compared to those solved in other space groups. Solvent molecules could be refined anisotropically in the crystal of Fe(bdp) at 100 K, accounting for all pore void space.

When the crystal was warmed to 298 K, the space group was determined to be *C*222<sub>1</sub> instead of *P*<sub>2</sub><sub>1</sub>/*c* and was refined as an inversion twin (batch scale factor, BASF = 0.52(4); the uncertainty corresponds to  $\pm 1$  s.d.). At 298 K, there was extensive solvent disorder that could not be modelled. A solvent mask was applied, as implemented in OLEX2, to account for unassigned electron density within the pores. The loss in intensity of spots upon warming to 298 K, and the large anisotropic displacement parameters that result from linker and solvent disorder, gave rise to A- and B-level alerts from checkCIF (<http://checkcif.iucr.org>). Responses addressing these alerts have been included in the CIF (crystallographic information file, available from the Cambridge Crystallographic Data Centre; see Author Information) and can be read in reports generated by checkCIF.

**Scanning electron microscopy.** Scanning electron microscopy (SEM) samples of Co(bdp) and Fe(bdp) were prepared by dispersing microcrystalline powders into DCM and drop casting onto a silicon chip (Extended Data Fig. 4). To dissipate charge, the samples were sputter coated with approximately 3 nm of Au (Denton Vacuum). Crystals were imaged at 5 keV and 12  $\mu$ A by field emission SEM (JEOL FSM6430).

**Microcalorimetry measurements.** Approximately 0.2 g of Co(bdp) was used for combined microcalorimetry and high-pressure CH<sub>4</sub> adsorption experiments. Before each experiment, samples were outgassed *ex situ* at 423 K for 16 h under a dynamic vacuum of  $10^{-3}$  mbar. The microcalorimetry experiments were performed using a custom-built manometric adsorption apparatus coupled with a Tian-Calvet-type microcalorimeter<sup>34</sup>. This experimental device allows the simultaneous determination of the adsorption isotherm and the adsorption enthalpy using a point-by-point introduction of gas to the sample. A multi-pneumovalve system allows the introduction of the adsorbate to the sample. An exothermic thermal effect accompanied each introduction, which is due to both the adsorption process and gas compression. This peak in the energy curve with time is thus integrated to calculate a pseudo-differential enthalpy of adsorption for each dose. Errors in this calculation can be estimated at  $\pm 1$  kJ mol<sup>-1</sup>. Experiments were carried out at 303 K and up to 70 bar with CH<sub>4</sub> of a purity of above 99.999%.

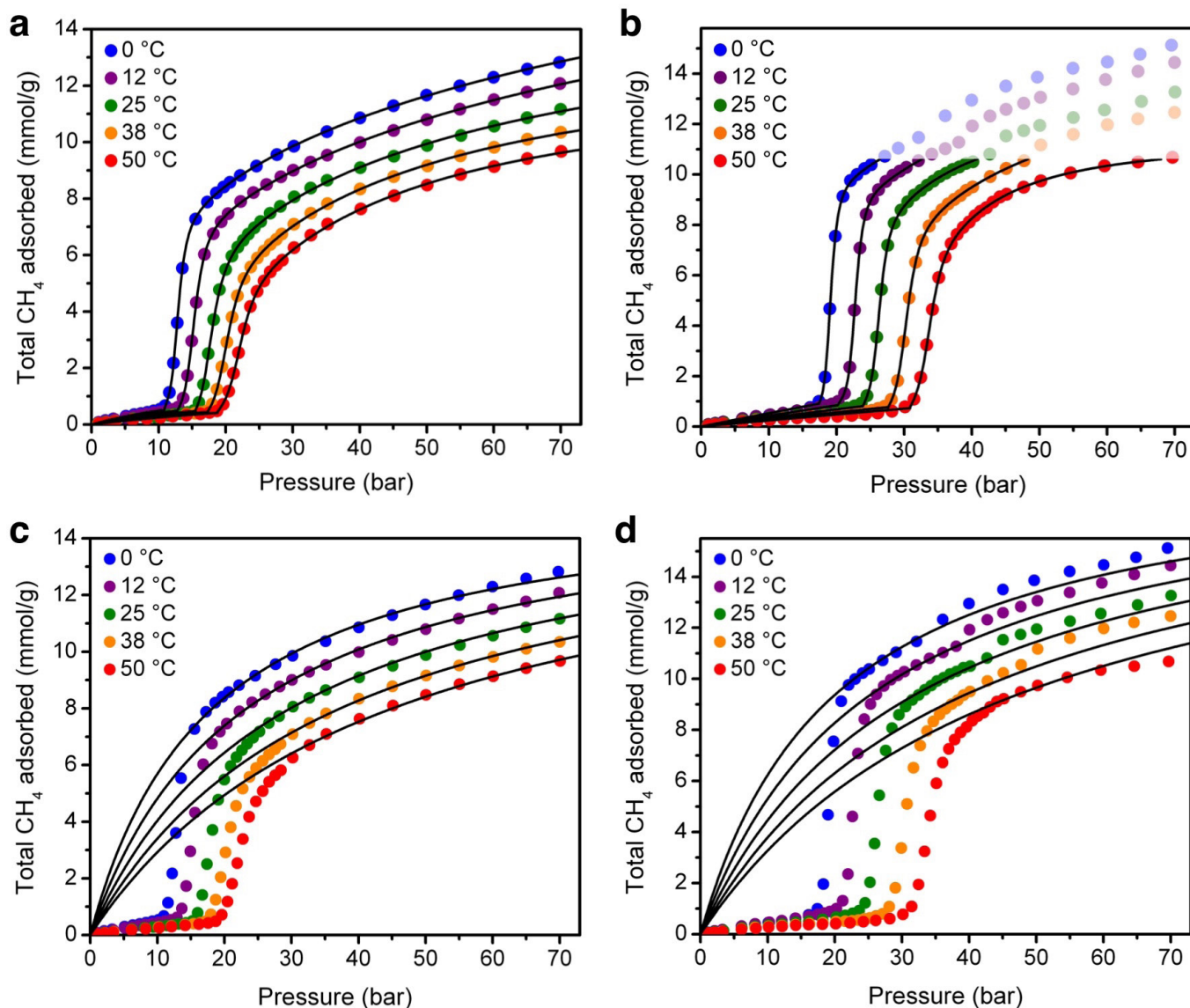
32. Lemmon, E. W., Huber, M. L. & McLinden, M. O. *NIST Standard Reference Database 23: Reference Fluid Thermodynamic and Transport Properties—REFPROP Version 8.0* (National Institute of Standards and Technology, 2007).
33. Lu, Y. *et al.* A cobalt(II)-containing metal-organic framework showing catalytic activity in oxidation reactions. *Z. Anorg. Allg. Chem.* **634**, 2411–2417 (2008).
34. Llewellyn, P. L. & Maurin, G. Gas adsorption microcalorimetry and modeling to characterise zeolites and related materials. *C. R. Chimie* **8**, 283–302 (2005).



**Extended Data Figure 1 | High-pressure  $\text{CH}_4$  cycling.** **a**, Excess  $\text{CH}_4$  isotherms at 25 °C for Co(bdp) repeated four times on the same sample, which was regenerated under vacuum at 25 °C for 2 h between measurements. The adsorption step is at a slightly higher pressure during the first run because there is probably a slightly higher energy barrier to the first expansion of a freshly packed sample; however, the desorption steps occur at identical pressures for all four runs. **b**, The adsorption and

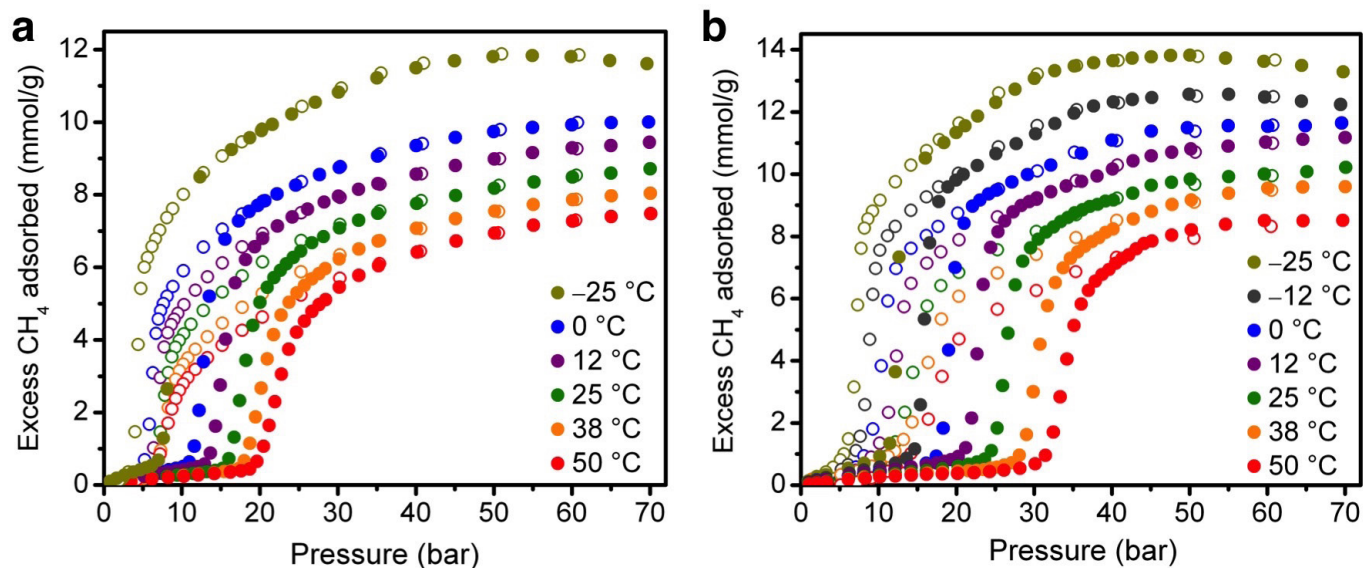
desorption pressures are shown as green and red circles, respectively, for 100  $\text{CH}_4$  adsorption–desorption cycles in Co(bdp) at 25 °C. **c**, Excess  $\text{CH}_4$  adsorption isotherms at 25 °C for Co(bdp) after 0, 25, 50, 75, and 100 cycles of 35-bar adsorption and 5-bar desorption. **d**, Excess  $\text{CH}_4$  isotherms at 25 °C for Co(bdp) before (green) and after (blue) the 100 adsorption–desorption cycles between 35 bar and 5 bar. Filled and open circles in **a** and **d** correspond to adsorption and desorption, respectively.



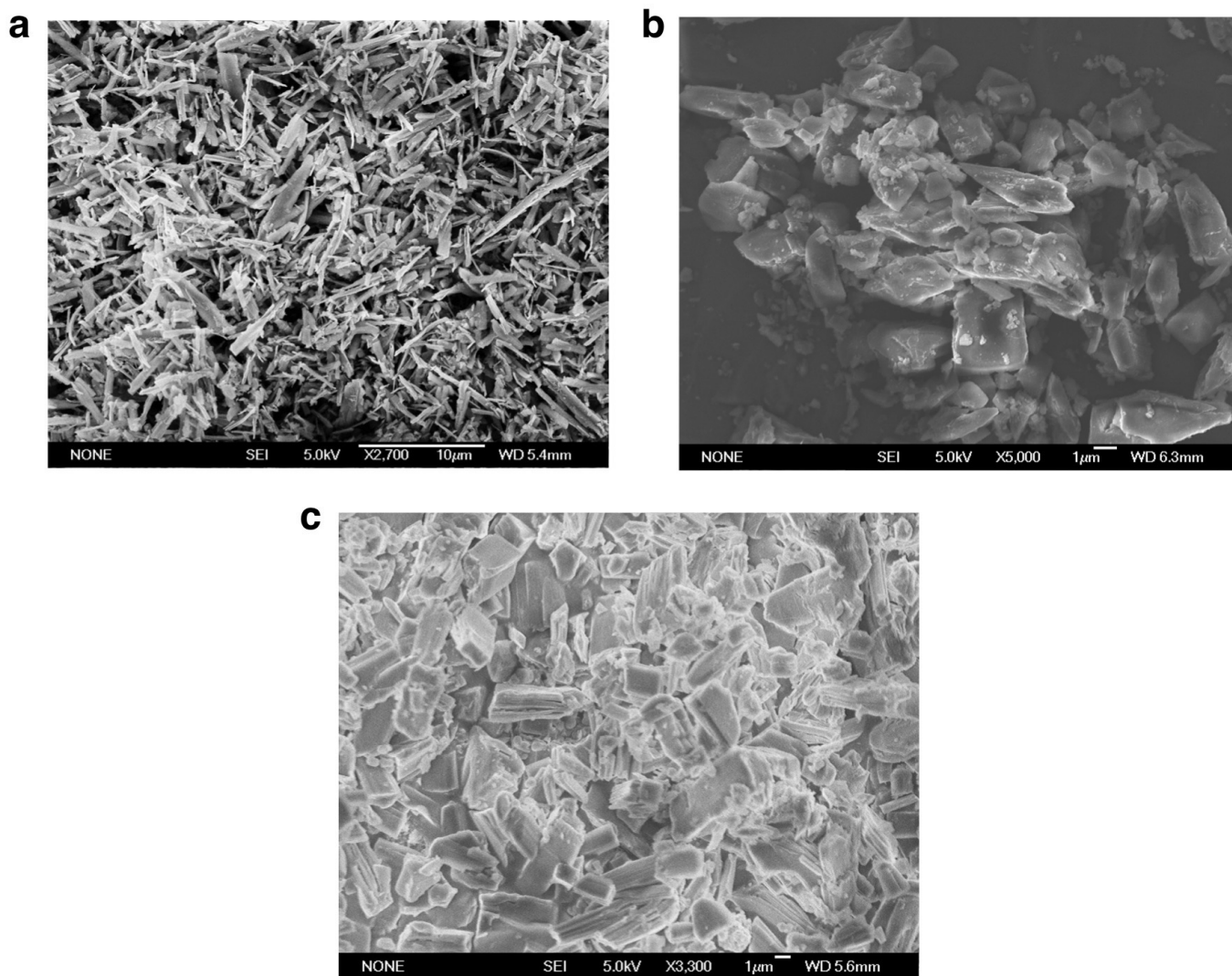


**Extended Data Figure 2 | Adsorption isotherm fitting.** **a**, Total CH<sub>4</sub> adsorption isotherms at 0 °C, 12 °C, 25 °C, 38 °C, and 50 °C for Co(bdp), with adsorption after the step fitted independently at each temperature with an offset dual-site Langmuir–Freundlich equation. The small pre-step adsorption was fitted with a single-site Langmuir model. **b**, Total CH<sub>4</sub> adsorption isotherms at 0 °C, 12 °C, 25 °C, 38 °C, and 50 °C for Fe(bdp) with adsorption after the phase transition fitted independently at each temperature with an offset dual-site Langmuir–Freundlich equation. The pre-step adsorption was fitted with a single-site Langmuir model, and the isotherms were only fitted to a maximum loading of 10.6 mmol g<sup>−1</sup>, as indicated by the shading, to avoid complications from the second transition

at higher CH<sub>4</sub> loadings. As such, differential enthalpies are only calculated up to a maximum loading of 10.6 mmol g<sup>−1</sup>. **c**, Total CH<sub>4</sub> adsorption isotherms at 0 °C, 12 °C, 25 °C, 38 °C, and 50 °C for Co(bdp) with the corresponding single-site Langmuir fit for CH<sub>4</sub> adsorption in the expanded phase. **d**, Total CH<sub>4</sub> adsorption isotherms at 0 °C, 12 °C, 25 °C, 38 °C, and 50 °C for Fe(bdp) with the corresponding single-site Langmuir fit for CH<sub>4</sub> adsorption in the 40-bar expanded phase. The data were only fitted for the region of the isotherms that falls after the initial hysteresis loop closes and before the second isotherm step. All single- and dual-site Langmuir–Freundlich fits are shown as black lines.

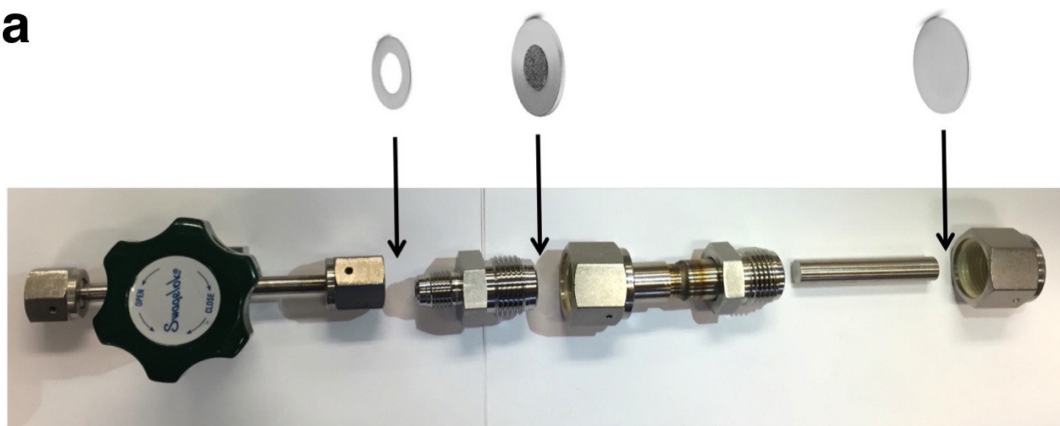
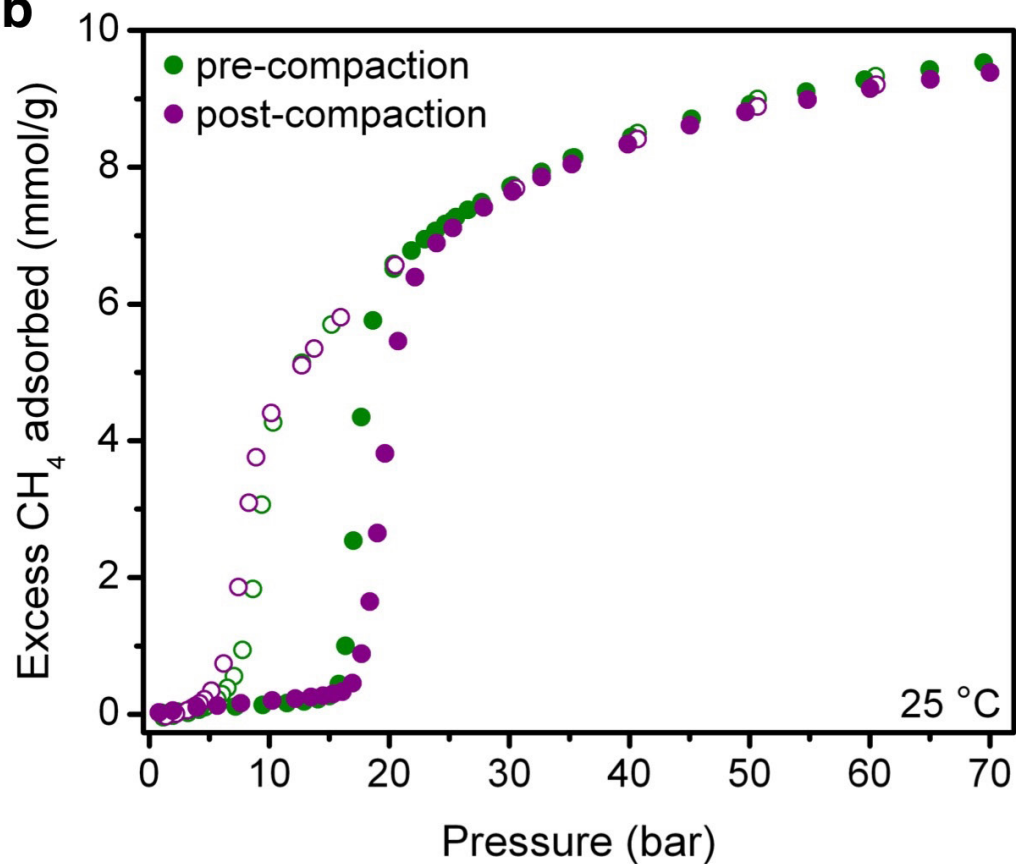


**Extended Data Figure 3 | Excess  $\text{CH}_4$  adsorption data.** **a**, Excess  $\text{CH}_4$  adsorption isotherms at  $-25^\circ\text{C}$ ,  $0^\circ\text{C}$ ,  $12^\circ\text{C}$ ,  $25^\circ\text{C}$ ,  $38^\circ\text{C}$ , and  $50^\circ\text{C}$  for  $\text{Co}(\text{bdp})$ . **b**, Excess  $\text{CH}_4$  adsorption isotherms at  $-25^\circ\text{C}$ ,  $-12^\circ\text{C}$ ,  $0^\circ\text{C}$ ,  $12^\circ\text{C}$ ,  $25^\circ\text{C}$ ,  $38^\circ\text{C}$ , and  $50^\circ\text{C}$  for  $\text{Fe}(\text{bdp})$ . Filled and open circles correspond to adsorption and desorption, respectively.



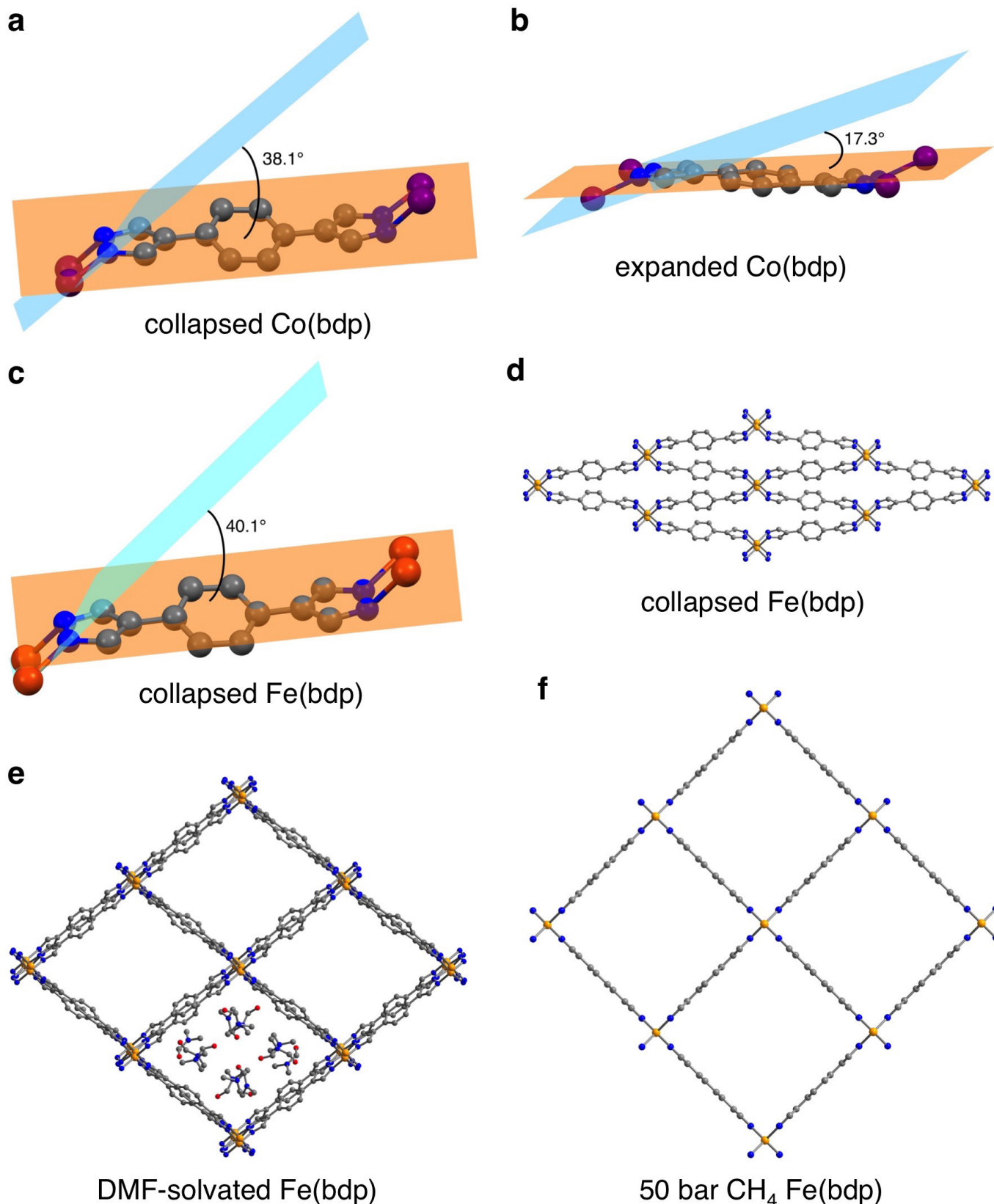
**Extended Data Figure 4 | SEM images.** **a**, SEM image of DMF-solvated Co(bdp) microcrystalline powder. Scale bar, 10  $\mu\text{m}$ . **b**, SEM image of Co(bdp) microcrystalline powder after more than 100 CH<sub>4</sub> adsorption-desorption cycles. Scale bar, 1  $\mu\text{m}$ . **c**, SEM image of desolvated Fe(bdp) microcrystalline powder. Scale bar, 1  $\mu\text{m}$ .



**a****b**

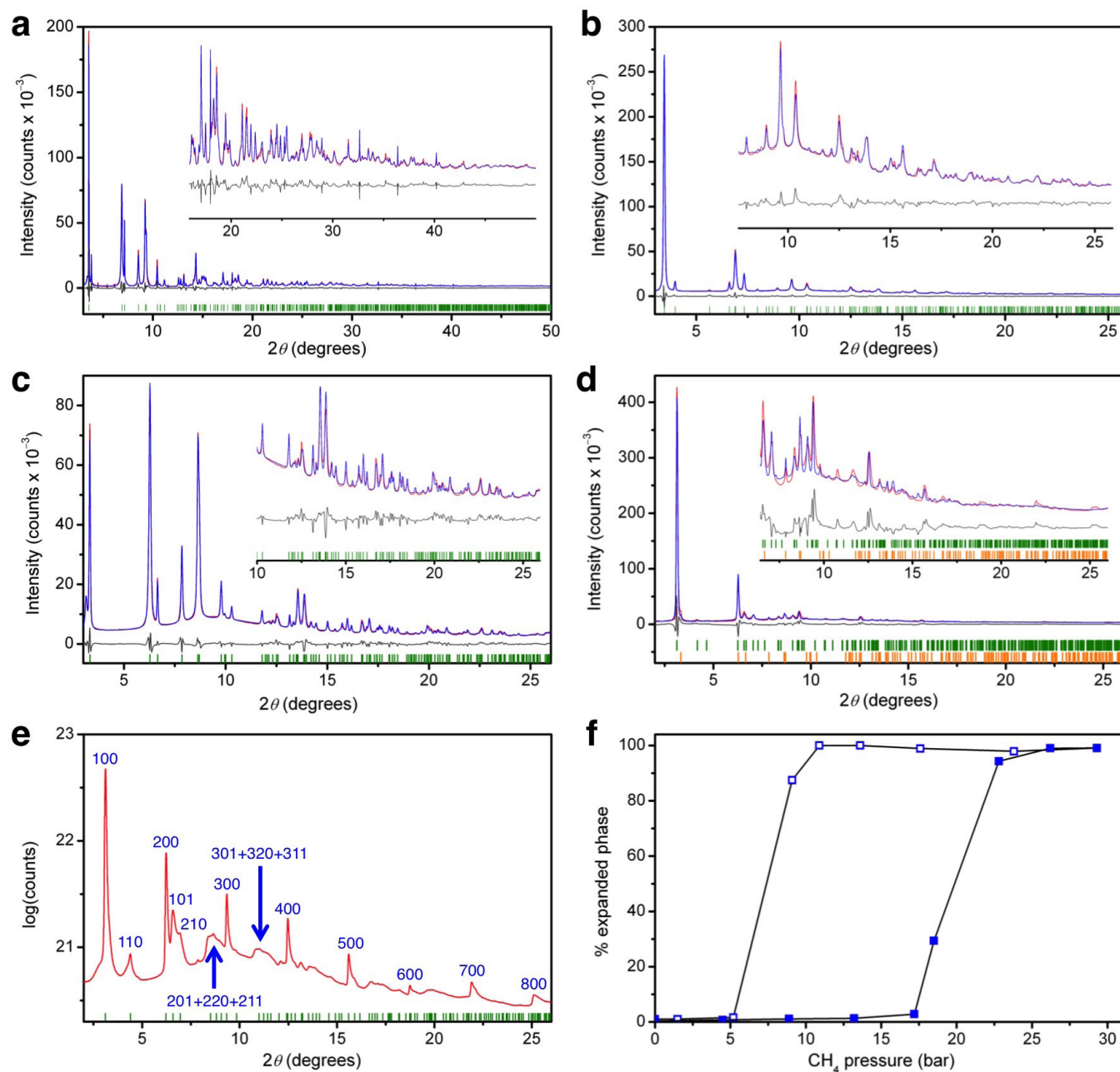
**Extended Data Figure 5 | Effect of mechanical pressure.** **a**, Sample holder used for combined applied mechanical pressure and high-pressure CH<sub>4</sub> adsorption experiments. The sample is located in the volume to the right of the fritted gasket and to the left of the blank gasket. A press is used to compact metal rods of different lengths against the sample, and the blank gasket is sealed behind the rod so that the uniaxial applied mechanical

pressure (and constricted volume) is maintained throughout the high-pressure CH<sub>4</sub> adsorption experiment. **b**, Excess CH<sub>4</sub> isotherms at 25 °C for Co(bdp) before (green) and after (purple) the applied mechanical pressure studies. Filled and open circles correspond to adsorption and desorption, respectively.



**Extended Data Figure 6 | Solid-state structures.** **a, b,** The angles between the plane of the pyrazolate (light orange) and the Co–N–N–Co plane (light blue) are 38.1° and 17.3° in the collapsed and the CH<sub>4</sub>-expanded phases of Co(bdp), respectively. **c,** The angle between the plane of the pyrazolate (light orange) and the Fe–N–N–Fe plane (light blue) is 40.1° in the collapsed phase

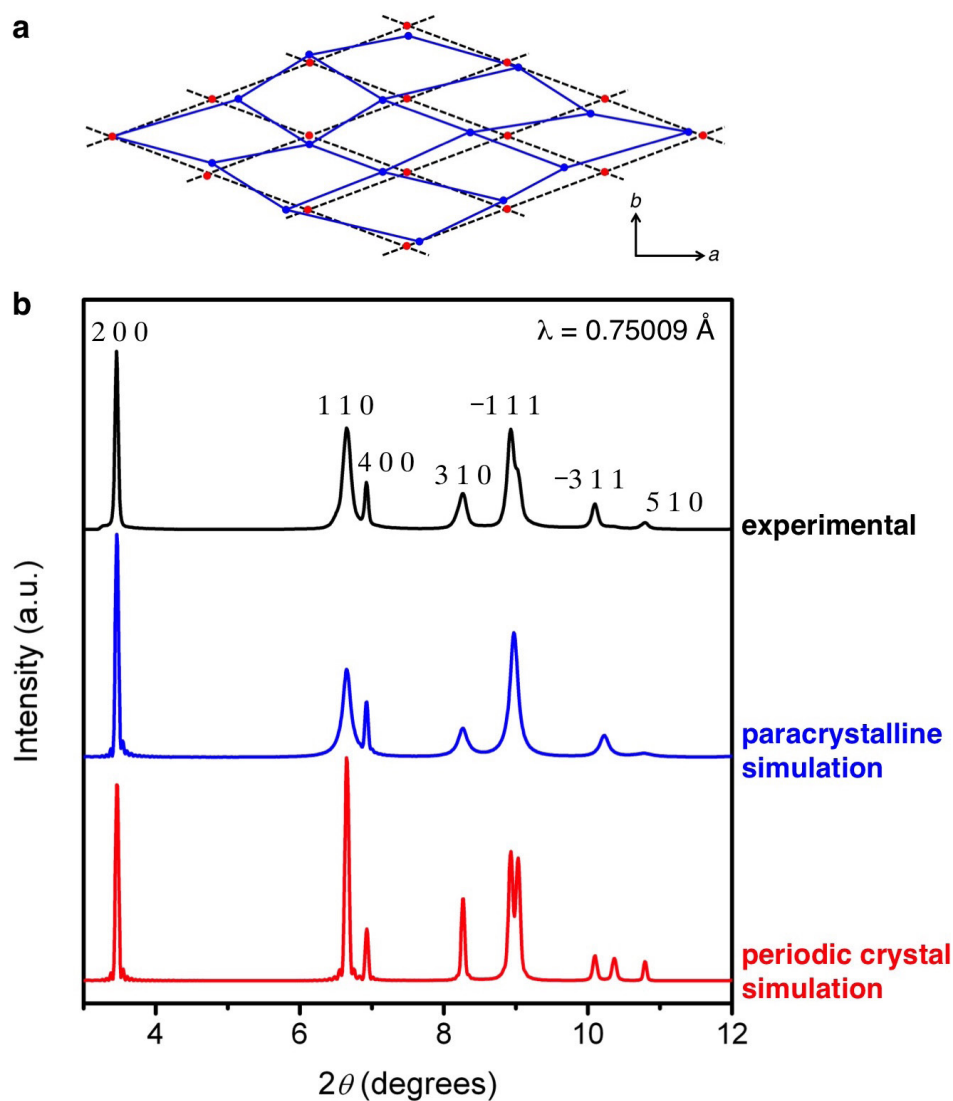
of Fe(bdp). **d,** Structure of the collapsed phase of Fe(bdp) under vacuum at 25 °C. **e,** Structure of the DMF-solvated phase of Fe(bdp) at 100 K. **f,** Idealized average structure of the 50-bar CH<sub>4</sub>-expanded phase of Fe(bdp) at 25 °C. In **a–f,** Grey, blue, red, purple, and orange spheres represent C, N, O, Co, and Fe atoms, respectively; H atoms are omitted for clarity.



**Extended Data Figure 7 | Powder X-ray diffraction.** **a–d**, Rietveld refinements for powder X-ray diffraction data ( $2\theta$  is the diffraction angle) for Co(bdp) at 25°C and under vacuum with  $\lambda = 0.77475$  Å (**a**), for Co(bdp) at 30 bar of CH<sub>4</sub> and 25°C with  $\lambda = 0.75009$  Å (**b**), for Fe(bdp) under vacuum at 25°C with  $\lambda = 0.72768$  Å (**c**), and for Fe(bdp) at 40 bar of CH<sub>4</sub> and 25°C with  $\lambda = 0.72768$  Å (**d**). Red and blue lines represent the observed and calculated diffraction patterns, respectively. Grey lines represent the difference between observed and calculated patterns, and green and orange tick marks indicate calculated Bragg peak positions. The broad hump observed at 10° in the diffraction patterns is due to diffuse

scattering from the sample holder (a thick-walled quartz glass capillary). The insets are magnified views of the main plots. **e**, Powder X-ray diffraction data for Fe(bdp) at 50 bar of CH<sub>4</sub> and 25°C ( $\lambda = 0.72768$  Å). Green tick marks indicate Bragg angles for space-group-permitted reflections; the corresponding Miller indices are indicated for the most prominent peaks. Blue arrows indicate broad humps where multiple reflections overlap. **f**, The percentage of the expanded phase of Co(bdp) that is present in the variable-pressure experimental powder X-ray diffraction patterns as a function of CH<sub>4</sub> pressure. The filled squares represent data collected during adsorption; the open squares represent data collected during desorption.





**Extended Data Figure 8 | Paracrystalline model.** **a**, An illustration of the paracrystalline distortion in the crystallographic  $a$ - $b$  plane of the collapsed phases of Co(bdp) and Fe(bdp) that leads to complex Bragg peak broadening. Black dashed lines represent the periodic crystal lattice; blue lines represent the paracrystal. Red circles represent the positions of metal-pyrazolate chains in the periodic lattice; blue circles represent their positions in a paracrystal. The magnitude of the paracrystalline distortion has been exaggerated for clarity. **b**, Simulated diffraction patterns are shown for a periodic collapsed Co(bdp) nanocrystal (75 nm  $\times$  60 nm  $\times$  43 nm; red trace)

and for a paracrystal of equivalent size (blue trace). The upper trace (black) corresponds to the background-subtracted experimental diffraction pattern of the collapsed phase of Co(bdp) at 25 °C; the corresponding Miller indices are indicated for the most prominent peaks. For clarity, the three patterns have been given an arbitrary  $y$  offset; a.u., arbitrary units. Similar anisotropic peak broadening, which inflates  $hk0$  peaks (but not  $h00$  or  $0k0$  ones), is clearly visible in the experimental diffraction pattern and the paracrystalline simulation. The exact full-widths at half-maximum for the experimental and simulated Bragg peaks are given in Supplementary Table 16.

# North Pacific deglacial hypoxic events linked to abrupt ocean warming

S. K. Praetorius<sup>1†</sup>, A. C. Mix<sup>1</sup>, M. H. Walczak<sup>1</sup>, M. D. Wolhowe<sup>1</sup>, J. A. Addison<sup>2</sup> & F. G. Prah<sup>1</sup>

Marine sediments from the North Pacific document two episodes of expansion and strengthening of the subsurface oxygen minimum zone (OMZ) accompanied by seafloor hypoxia during the last deglacial transition<sup>1–4</sup>. The mechanisms driving this hypoxia remain under debate<sup>1–11</sup>. We present a new high-resolution alkenone palaeotemperature reconstruction from the Gulf of Alaska that reveals two abrupt warming events of 4–5 degrees Celsius at the onset of the Bølling and Holocene intervals that coincide with sudden shifts to hypoxia at intermediate depths. The presence of diatomaceous laminations and hypoxia-tolerant benthic foraminiferal species, peaks in redox-sensitive trace metals<sup>12,13</sup>, and enhanced <sup>15</sup>N/<sup>14</sup>N ratio of organic matter<sup>13</sup>, collectively suggest association with high export production. A decrease in <sup>18</sup>O/<sup>16</sup>O values of benthic foraminifera accompanying the most severe deoxygenation event indicates subsurface warming of up to about 2 degrees Celsius. We infer that abrupt warming triggered expansion of the North Pacific OMZ through reduced oxygen solubility and increased marine productivity via physiological effects; following initiation of hypoxia, remobilization of iron from hypoxic sediments could have provided a positive feedback on ocean deoxygenation through increased nutrient utilization and carbon export. Such a biogeochemical amplification process implies high sensitivity of OMZ expansion to warming.

Models suggest enhanced ocean deoxygenation in response to future global warming, owing to both a reduction in oxygen solubility and decreased subsurface ventilation related to thermal stratification<sup>14,15</sup>. Uncertainty in these projections reflects weak constraints on the response of marine primary productivity to warming<sup>16</sup> and the extent to which ecosystem changes will translate to carbon export and remineralization<sup>17</sup>, thus altering subsurface oxygen demand. Once hypoxia is initiated, seafloor biogeochemical cycling may sustain and enhance low-O<sub>2</sub> conditions via a threshold effect<sup>18</sup>. For example, reductive mobilization of sedimentary iron may further stimulate primary productivity in surface waters of high-nitrate-low-chlorophyll (HNLC) regions if it occurs within an ocean depth range susceptible to mixing into the euphotic zone; this in turn would increase oxygen demand in underlying waters, until sulfidic conditions ensue and limit further supplies of dissolved iron from the sediments to the upper water column<sup>19</sup>.

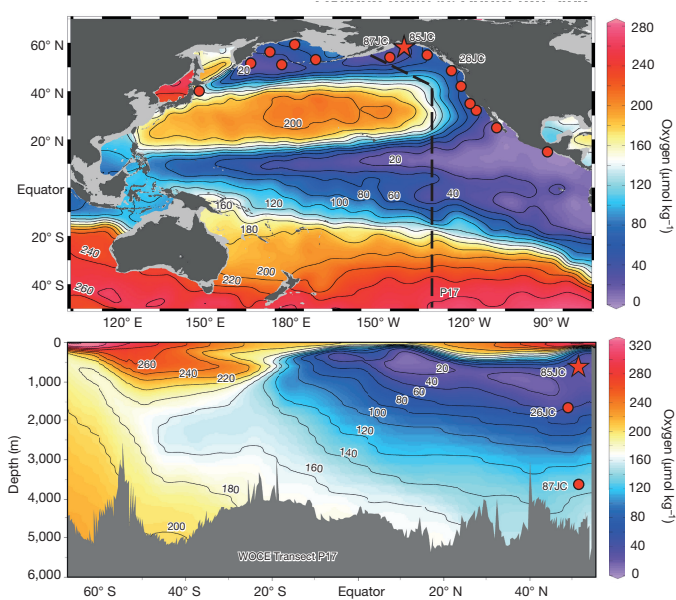
Data from intermediate water depth marine sediment cores across the North Pacific (Fig. 1) document expansion and strengthening of the OMZ during the Bølling–Allerød interstade (14.7–12.9 thousand years ago (ka)) and earliest part of the Holocene interglacial (11.5–10.5 ka)<sup>1–4</sup>. Mechanisms proposed to account for deglacial hypoxia include a decrease in ocean ventilation related to changes in ocean circulation<sup>3,5</sup>, and increased oxygen demand related to enhanced export productivity<sup>6–11</sup>.

So far, no consistent evidence links decreased ventilation rate with these hypoxic events. Benthic-planktonic radiocarbon age differences on the Gulf of Alaska margin show no significant deviation from the

long-term mean ( $725 \pm 200$  yr) during either of the hypoxic events (Fig. 2), and when benthic-planktonic increases do occur (during the Younger Dryas interval; 12.9–11.7 ka), they are not associated with hypoxia<sup>20</sup>.

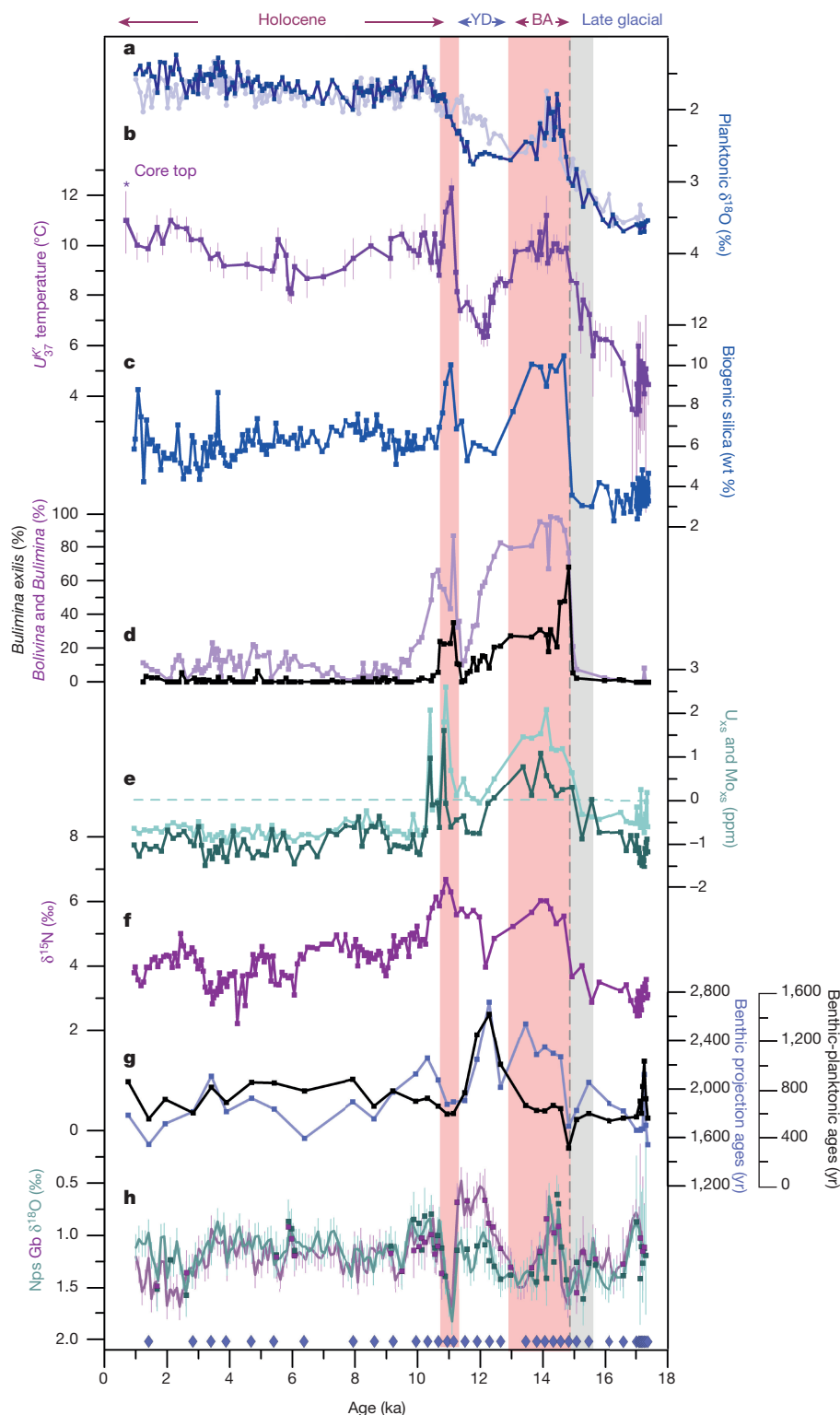
Mechanisms to stimulate productivity over such widespread oceanographic settings remain unclear. However, once hypoxia is initiated in a subsurface water mass, it may spread over a greater area than the region of elevated productivity<sup>6</sup>, because downward mixing of oxygen beyond the range of shallow wind-driven mixing is small. Iron release from continental shelves in response to sea-level rise has been suggested as a potential driver of increased productivity in HNLC regions such as the subpolar North Pacific<sup>1,2</sup>. However, regional sea-level histories differ owing to isostatic responses to ice unloading. For the sea-level mechanism to work, long-distance transport of iron would be required.

Alternative hypothetical mechanisms to stimulate North Pacific productivity have invoked increased upwelling of macronutrients during hypoxic events<sup>9</sup>. This would require reduced stratification, paradoxically at a time of northern hemisphere warming and enhanced freshwater input to the marginal ocean from melting ice sheets. This mechanism does not account for the fact that macronutrients such



**Figure 1 | Study area and core locations.** Top, colour shading on map shows modern oxygen concentration at 400 m depth, with the clearly defined OMZ across the North, Eastern, and Equatorial Pacific. The core site EW0408-85JC in this study is indicated with a red star; other core sites that document deglacial hypoxic/productivity events are indicated with red circles<sup>4</sup>. Bottom, meridional cross-section of oxygen concentration in the modern Eastern Pacific, with the location of core sites used in the  $\delta^{18}\text{O}$  depth transect (Fig. 4). Maps were generated with Ocean Data View<sup>31</sup>.

<sup>1</sup>College of Earth, Ocean, and Atmospheric Sciences, Oregon State University, Corvallis, Oregon 97331, USA. <sup>2</sup>US Geological Survey, Menlo Park, California 94025, USA. <sup>†</sup>Present address: Department of Global Ecology, Carnegie Institution for Science, Stanford, California 94305, USA.



**Figure 2 | Data from core EW0408-85JC.** **a**, Planktonic  $\delta^{18}\text{O}$  data (Nps: dark blue, Gb: light blue)<sup>2</sup>. **b**, Alkenone palaeotemperature (purple) and error estimates (light purple bars). **c**, Biogenic opal percentages relative to bulk sediment (blue)<sup>2</sup>. **d**, Relative abundance of low-oxygen benthic foraminifera (*Bolivina* and *Bulimina* genera: light purple, *Bulimina exilis*: black). **e**, Redox-sensitive trace metal concentrations (Mo is light green, U is dark green) in excess values relative to lithogenic background (dashed line)<sup>12,13</sup>. **f**, Sedimentary  $\delta^{15}\text{N}$  (dark violet)<sup>13</sup>. **g**, Benthic-planktonic radiocarbon age differences (black) and benthic projection ages calculated with respect to atmospheric  $\Delta^{14}\text{C}$  (blue)<sup>20</sup>. **h**, Reconstructions of near-surface seawater  $\delta^{18}\text{O}$  based on the planktonic species Gb (green) and

Nps (light violet), with depth-based pairs indicated as square symbols and values linearly interpolated at 100-yr intervals shown as a trend line. Age controls are from Davies-Walczak *et al.*<sup>20</sup> and are indicated with blue diamonds at the bottom of the plot. The pink bars represent the two laminated intervals in core EW0408-85JC, the grey bar indicates the zone in which changes in SST, trace metals, and benthic fauna slightly precede the onset of laminations (dashed grey line) (expanded view in Extended Data Fig. 7). The timing of major climate intervals are indicated at the top of the plot: Holocene (11.6–0 ka), Younger Dryas (YD; 12.9–11.7 ka), Bölling–Allerød (BA; 14.6–12.9 ka), late glacial (18–14.7 ka). Nps, *Neogloboquadrina pachyderma*; Gb, *Globigerina bulloides*.



as phosphate and nitrate, supplied today by vertical mixing, do not limit productivity in the region. Further, most deep mixing occurs in winter, when light is limiting. Although deep upwelling also provides some iron, it is not currently enough to consume the macronutrients; relief of iron limitation in the past would require an additional iron source. Haline stratification due to ice melt may have enhanced marine productivity by reducing deep mixing of plankton out of the euphotic zone<sup>10</sup>. Haline inhibition of deep mixing would, however, reduce the source of subsurface iron<sup>21</sup>.

Distinguishing among the various hypotheses requires separating sea-surface temperature (SST) and salinity effects on near-surface stratification and subsurface ventilation. Here we pair a new high-resolution palaeotemperature record based on the  $U_{37}^K$  alkenone index from the Gulf of Alaska with benthic faunal assemblages and proxies for ventilation rate, export productivity, and surface stratification to evaluate the sequence of oceanographic changes leading to hypoxia in a marine sediment core located in the upper reaches of the modern OMZ (EW0408-85JC, 59° 33.32' N, 144° 9.21' W, 682 m depth). The various proxy indicators are co-registered with little or no impact from chronologic error, because they occur in the same sediment core. We find that two abrupt deglacial warming events of 4–5 °C coincide with increases in nutrient utilization, export productivity, and the sudden onset of hypoxia. Alkenone SST reconstructions from other sites in the North Pacific show similar trends (Extended Data Fig. 9), indicating these were significant regional temperature fluctuations.

Palaeotemperatures in the northern Gulf of Alaska were lowest (~5 °C) near 17.0 ka, (Fig. 2), coincident with a peak in ice-rafted debris (IRD)<sup>2</sup>. Warming commenced near 16.5 ka, before the Bølling interstade in Greenland. Warming then accelerated, with a rapid 3–4 °C rise from 15.2 to 14.7 ka, coincident with the Bølling onset; warm conditions persisted until 13.0 ka.

The high relative abundance of the benthic foraminifera *Epistominella pacifica* during the late glacial period (17–15 ka) indicates that the water column was less oxygenated than modern conditions (Extended Data Fig. 4). Following the accelerated warming into the Bølling, severe hypoxia seems to have started abruptly at 14.7 ka, as documented by a sharp transition to sediment laminations, a shift to benthic populations dominated by the low-oxygen foraminifera *Bulimina exilis*, and an increase in sedimentary molybdenum and uranium concentrations (Fig. 2). This transition coincided with increased concentrations of biogenic silica and marine organic carbon and a rise in the  $\delta^{15}\text{N}$  of organic matter, reflecting enhanced nutrient utilization and export productivity<sup>2,13</sup> (Extended Data Figs 5 and 6).

The increase in redox-sensitive trace metals (excess Mo and U)<sup>12,13</sup> and low-oxygen benthic species slightly preceded the increase in biogenic silica and total organic carbon (TOC) (by 2–5 cm, about 300 ± 100 yr) during the Bølling–Allerød transition (Extended Data Fig. 7). This implies that local hypoxia may have developed before the increase in export productivity, most likely through a reduction in oxygen solubility related to ocean warming. In the absence of bioturbation, cm-scale depth offsets in these proxies may in part reflect redox gradients in the sediment column. However, a deeper site in the Gulf of Alaska (EW0408-26JC; 1,620 m) shows benthic faunal evidence for a gradual decrease in oxygen, starting with the early warming at 16 ka (Extended Data Fig. 8), well in advance of productivity enhancement.

The benthic faunal assemblages and trace metals indicate that hypoxic conditions persisted while surface palaeotemperatures remained elevated near 10 °C during the Bølling–Allerød interstade. More oxygenated conditions returned only when alkenone palaeotemperatures fell by 4 °C during the Younger Dryas interval (12.9–11.7 ka). Another abrupt 5 °C warming occurred during the transition into the early Holocene (11.4–10.9 ka); SSTs above 11 °C were again accompanied by an increase in biogenic silica concentrations, a rise in  $\delta^{15}\text{N}$ , and a return to hypoxic conditions (evidenced by laminations, low-oxygen benthic fauna, and an increase in authigenic Mo and U concentrations).

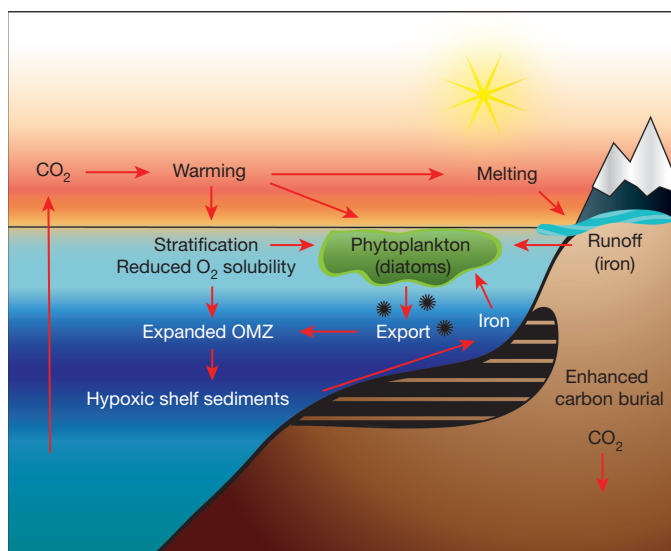
Benthic faunal assemblage data document crossing of an ecological threshold during both episodes of hypoxia, with an abrupt transition from virtual absence to dominance of *Bulimina exilis* at the start of each event, indicating that the system switched nearly instantaneously from intermediate to severe hypoxia. Shortly following the initial peaks in *Bulimina exilis*, the site was re-colonized by other low-oxygen *Bolivina* species (Fig. 2), indicating a slight relaxation from extreme to strong hypoxia. *Epistominella pacifica* reappears in the faunal assemblages during the Younger Dryas and early Holocene, consistent with the return of more oxic conditions in bottom waters as the surface ocean cooled (Extended Data Fig. 4). Some low-oxygen benthic fauna and redox-sensitive trace metals persisted after decreases in SST and reduction in biogenic silica, suggesting a lag in the amelioration of hypoxic conditions. A trend of increasing relative percentage of *Uvigerina peregrina* during the Holocene indicates the water column became progressively well oxygenated.

The alkenone palaeotemperature data show that most of the anomalously low  $\delta^{18}\text{O}$  at onset of the Bølling interstade (1.1‰ of the total 1.5‰ change) reflects warming (Fig. 2). The temperature and ice-volume-corrected seawater  $\delta^{18}\text{O}$  record is primarily an indicator of freshwater input from land, which is typically high in this stormy and glaciated region. Freshwater input varies near the onset of the Bølling–Allerød interstade hypoxic event, but there are no consistent trends that would indicate a sustained freshening of the surface ocean during either hypoxic event. This finding precludes haline stratification as a primary cause of hypoxia at this location. Accumulation of terrigenous silt is also low during the hypoxic intervals, arguing against glacial runoff and rock flour as a primary source of iron to fertilize high productivity at these times. However, high rates of terrigenous sediment accumulation on the shelf and slope during the late glacial period<sup>2</sup> would have provided a reservoir of iron in the sediments, which may then have been available for release in a bioavailable form upon initiation of hypoxia.

Warming slightly precedes both the increase in productivity and the onset of hypoxia (rates of warming are 0.6–1.2 °C per century in the lead up to both events; Extended Data Fig. 3), making it likely that temperature exerted a primary initial trigger for biogeochemical amplifying effects responsible for extensive hypoxia. This observation narrows the hypotheses to two (Fig. 3): either (1) rapid warming led to modest regional hypoxia via thermal solubility effects, which in turn stimulated marine primary productivity indirectly through the mobilization of iron from hypoxic sediments, or (2) rapid warming directly stimulated marine productivity, which led to higher consumption of oxygen in subsurface waters through remineralization of organic matter, with possible further amplification via reductive iron release. In both hypotheses, warming leads to hypoxia; they are not mutually exclusive. Both suggest increases in nutrient utilization in what is now an iron-limited system; this is supported by rising  $\delta^{15}\text{N}$  of organic matter<sup>13</sup> (Fig. 2) coupled to an increase in  $\delta^{13}\text{C}$  in planktonic foraminifera (Extended Data Figs 5 and 6) during the warm events.

The first hypothesis requires warming in subsurface waters. If the  $\delta^{18}\text{O}$  excursion (–1‰) in the benthic record were due to an increase in temperature alone, it would imply subsurface warming of about 4–5 °C at a depth of 682 m. The presumed implausibility of such warming at depth has led to the interpretation of this feature as a pulse of low salinity waters<sup>2</sup>. However, the occurrence of similar but smaller benthic  $\delta^{18}\text{O}$  anomalies (–0.4–0.6‰) in deeper sites during the Bølling–Allerød interstade (Fig. 4, Extended Data Fig. 1) is consistent with up to about 2.0 °C abyssal warming; pulses of low salinity are implausible in the deep basin.

Based on known thermal solubility effects, a 2.0 °C warming would reduce subsurface oxygen concentrations by about 17  $\mu\text{mol kg}^{-1}$  (ref. 22), which if initiated relative to the modern OMZ of the North Pacific, would drive the site EW0408-85JC to less than 5  $\mu\text{mol kg}^{-1}$ , enough for significant ecological impacts<sup>23</sup> (Fig. 1). The benthic fauna

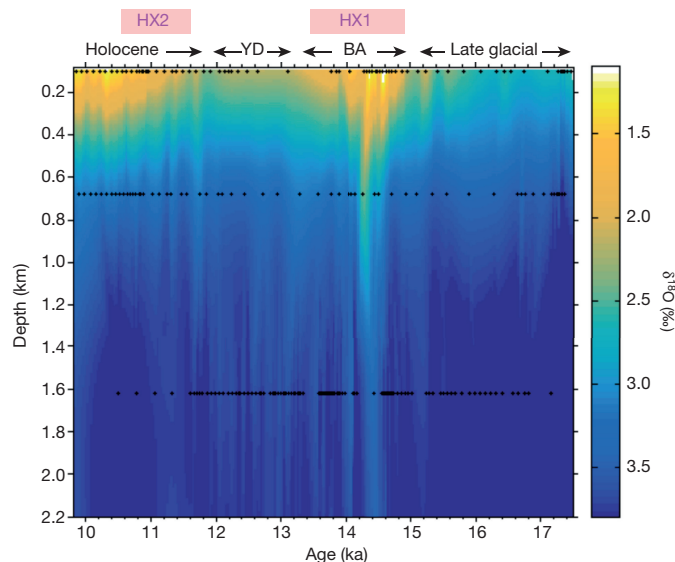


**Figure 3 | Schematic diagram of feedback processes linking ocean warming to enhanced export productivity.** As CO<sub>2</sub> is vented from the deep ocean throughout the deglaciation, the ocean warms along with the climate. Warming reduces oxygen solubility and promotes stratification, leading to an expansion of the OMZ and a greater area of suboxic sediments in the shallow subsurface, which remobilizes bioavailable iron to fuel marine productivity, further depleting subsurface oxygen concentrations as organic matter is exported and respired at depth. The availability of iron and warm, stratified conditions favours large diatoms<sup>27</sup>, which have high export efficiency owing to high settling velocities.

suggest slightly lower oxygenation than today before the Bølling–Allerød interstade (Extended Data Fig. 4), so subsurface thermal solubility effects might have been sufficient to trigger local hypoxia and initial vertical expansion of the OMZ, but are unlikely to drive widespread or severe hypoxia. If shoaling of the hypoxic boundary layer was sufficient to trigger an iron and phosphate<sup>24</sup> release from shallow sediments (that is, within reach of mixing into the euphotic zone), an ecological response favouring carbon export would have amplified initial thermally driven hypoxia. Production of larger diatoms enhances carbon export to the subsurface<sup>25</sup>, and these diatom species are known to increase in relative abundance in response to increased iron input<sup>26</sup> and stratification<sup>27</sup>. Dominance of large diatoms in the overall species assemblage has been found elsewhere in the Northeast Pacific during deglacial warming<sup>17</sup>.

The second hypothesis revives a longstanding debate in biological oceanography following Eppley's inference<sup>28</sup> that warming enhances phytoplankton growth rate directly through a Q<sub>10</sub> (exponential physiological rate) effect. Such an effect may be important in subpolar regions, and may be enhanced by stratification that keeps phytoplankton near the well-illuminated, warmer sea surface and is classically associated with high nutrient uptake and export<sup>29</sup>. The Eppley hypothesis remains controversial. Thermal Q<sub>10</sub> effects on productivity, although perhaps real at high latitudes, are not supported at low latitudes<sup>16</sup>, making it an unlikely mechanism to account for the inferred increase in deglacial productivity in lower latitude regions<sup>7</sup>.

Subsurface warming would increase remineralization rates of organic matter sinking out of the near-surface ocean, and thus biological oxygen demand in the zone most sensitive to hypoxia. As with the first scenario, carbon export to the subsurface ocean, rather than primary productivity itself, would be the key variable, suggesting control by ecosystem effects that favour large diatoms (such as iron availability), or other effects on particle sinking rates. Thus both scenarios imply thermal triggers, but both require biogeochemical amplification to sustain subsurface hypoxia, plausibly through reductive iron remobilization on continental margins<sup>19</sup>. Such a feedback between ocean warming, OMZ



**Figure 4 | Deglacial depth transect of  $\delta^{18}\text{O}$  in the Gulf of Alaska.** Planktonic foraminiferal data (Nps  $\delta^{18}\text{O}$ ) from core EW0408-85JC is used for the subsurface thermocline (depth = 100 m). Benthic foraminiferal  $\delta^{18}\text{O}$  data (expressed as *Uvigerina peregrina* equivalent values) from various core sites are plotted at depth: EW0408-85JC (682 m)<sup>2</sup>, EW0408-26JC (1,620 m), and EW0408-87JC (3,680 m) (data plotted in Extended Data Fig. 1). The plot is truncated at 2.2 km depth for an expanded view of intermediate depths. All data are corrected for the global isotopic effects of changing ice volume<sup>37</sup>. Labels identify climate intervals specified in Fig. 2. The timing of the deglacial hypoxic events are indicated with pink bars. Details of the age models of individual cores are included in Methods.

expansion, and marine productivity may also provide a plausible mechanism to explain links between interstadial warm periods and abrupt transitions to hypoxia observed throughout the North Pacific<sup>3,4,7,11</sup>, without the need to invoke changes in the Atlantic meridional overturning circulation and associated multi-century time lags of nutricline adjustment within the global ocean<sup>8</sup>.

Projected future warming of the subpolar North Pacific will probably exceed the temperatures associated with past hypoxic events by the mid-twenty-first century<sup>30</sup>, at sustained rates comparable to those preceding the deglacial hypoxic events. If enhanced biological productivity amplifies future deoxygenation as our evidence suggests it has in the past, substantial expansion of subsurface hypoxia beyond that predicted solely from thermal solubility effects may occur. While severe hypoxia would be catastrophic in the near-term for marine ecosystems and fisheries<sup>18</sup>, the resulting reduction of carbon remineralization rates and enhanced burial of organic matter associated with hypoxia may also provide a long-term negative feedback on rising CO<sub>2</sub> and greenhouse-driven warming, as may have occurred during the deglacial hypoxic events<sup>4</sup> (Extended Data Fig. 9).

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 9 February; accepted 17 September 2015.**

1. Mix, A. C. *et al.* Rapid climate oscillations in the Northeast Pacific during the last deglaciation reflect Northern and Southern Hemisphere sources, in *Mechanisms of global climate change at millennial time scales*, American Geophysical Union, edited by P. U. Clark *et al.*, Geophysical Monograph **112**, 127–148 (1999).
2. Davies, M. H. *et al.* The deglacial transition on the southeastern Alaska Margin: Meltwater input, sea level rise, marine productivity, and sedimentary anoxia. *Paleoceanography* **26**, PA2223 (2011).
3. Behl, R. J. & Kennett, J. P. Brief interstadial events in the Santa Barbara basin, NE Pacific, during the past 60 kyr. *Nature* **379**, 243–246 (1996).
4. Jaccard, S. L. & Galbraith, E. D. Large climate-driven changes in oceanic oxygen concentrations during the last deglaciation. *Nature Geosci.* **5**, 151–156 (2012).

5. Okazaki, Y. *et al.* Deepwater formation in the North Pacific during the Last Glacial Termination. *Science* **329**, 200–204 (2010).
6. Crusius, J., Pedersen, T. F., Kienast, S., Keigwin, L. & Labeyrie, L. Influence of northwest Pacific productivity on North Pacific Intermediate Water oxygen concentrations during the Bølling-Allerød interval (14.7–12.9 ka). *Geology* **32**, 633–636 (2004).
7. Hendy, I. L., Pedersen, T. F., Kennett, J. P. & Tada, R. Intermittent existence of a southern Californian upwelling cell during submillennial climate change of the last 60 kyr. *Paleoceanography* **19**, PA3007 (2004).
8. Schmittner, A., Galbraith, E. D., Hostetler, S. W., Pedersen, T. F. & Zang, R. Large fluctuations of dissolved oxygen in the Indian and Pacific oceans during Dansgaard-Oeschger oscillations caused by variations of North Atlantic Deep Water subduction. *Paleoceanography* **22**, PA3207 (2007).
9. Kohfeld, K. E. & Chase, Z. Controls on deglacial changes in biogenic fluxes in the North Pacific ocean. *Quat. Sci. Rev.* **30**, 3350–3363 (2011).
10. Lam, P. J. *et al.* Transient stratification as the cause of the North Pacific productivity spike during deglaciation. *Nat. Geosci.* **6**, 622–626 (2013).
11. Kuehn, H. *et al.* Laminated sediments in the Bering Sea reveal atmospheric teleconnections to Greenland climate on millennial to decadal timescales during the last deglaciation. *Clim. Past* **10**, 2215–2236 (2014).
12. Barron, J. A., Bukry, D., Dean, W. E., Addison, J. A. & Finney, B. Paleoceanography of the Gulf of Alaska during the past 15,000 years: results from diatoms, silicoflagellates, and geochemistry. *Mar. Micropaleontol.* **72**, 176–195 (2009).
13. Addison, J. A. *et al.* Productivity and sedimentary  $\delta^{15}\text{N}$  variability for the last 17,000 years along the northern Gulf of Alaska slope. *Paleoceanography* **27**, PA1206 (2012).
14. Keeling, R. F., Kortzinger, A. & Gruber, N. Ocean deoxygenation in a warming world. *Annu. Rev. Mar. Sci.* **2**, 199–229 (2010).
15. Schmittner, A., Oschlies, A., Matthews, H. D. & Galbraith, E. D. Future changes in climate, ocean circulation, ecosystems and biogeochemical cycling simulated for a business-as-usual  $\text{CO}_2$  emission scenario until year 4000 AD. *Glob. Biogeochem. Cycles* **22**, GB1013 (2008).
16. Behrenfeld, M. J. *et al.* Climate-driven trends in contemporary ocean productivity. *Nature* **444**, 752–755 (2006).
17. Lopes, C., Kucera, M. & Mix, A. C. Climate change decouples oceanic primary and export productivity and organic carbon burial. *Proc. Natl Acad. Sci. USA* **112**, 332–335 (2014).
18. Diaz, R. J. & Rosenberg, R. Spreading dead zone and consequences for marine ecosystems. *Science* **321**, 926–929 (2008).
19. Scholz, F., McManus, J., Mix, A. C., Hensen, C. & Schneider, R. R. The impact of ocean deoxygenation on iron release from continental margin sediments. *Nat. Geosci.* **7**, 433–437 (2014).
20. Davies-Walczak, M. H. *et al.* Late glacial to Holocene radiocarbon constraints on North Pacific Intermediate Water ventilation and deglacial atmospheric  $\text{CO}_2$  sources. *Earth Planet. Sci. Lett.* **397**, 57–66 (2014).
21. Takeda, S. Iron and phytoplankton growth in the subarctic North Pacific. *Aqua-BioScience Monographs* **4**, 41–93 (2011).
22. Benson, B. B. & Krause, D. J. The concentration and isotopic fractionation of oxygen dissolved in freshwater and seawater in equilibrium with the atmosphere. *Limnol. Oceanogr.* **29**, 620–632 (1984).
23. Hofmann, A. F. *et al.* Hypoxia by degrees: establishing definitions for a changing ocean. *Deep Sea Res. Part I Oceanogr. Res. Pap.* **58**, 1212–1226 (2011).
24. Ingall, E. & Jahnke, R. Evidence for enhanced phosphorus regeneration from marine sediments overlain by oxygen depleted waters. *Geochim. Cosmochim. Acta* **58**, 2571–2575 (1994).
25. Boyd, P. & Newton, P. Evidence of the potential influence of planktonic community structure on the interannual variability of particulate organic carbon flux. *Deep Sea Res. Part I Oceanogr. Res. Pap.* **42**, 619–639 (1995).
26. Hoffmann, L. J., Peeken, I., Lochte, K., Assmy, P. & Veldhuis, M. Different reaction of Southern Ocean phytoplankton size classes to iron fertilization. *Limnol. Oceanogr.* **51**, 1217–1229 (2006).
27. Kemp, A. E. S. & Villareal, T. A. High diatom production and export in stratified waters – A potential negative feedback to global warming. *Prog. Oceanogr.* **119**, 4–23 (2013).
28. Eppley, R. W. Temperature and phytoplankton growth in the sea. *Fish Bull.* **70**, 1063–1085 (1972).
29. Sverdrup, H. U. On conditions for the vernal blooming of phytoplankton. *ICES J. Mar. Sci.* **18**, 287–295 (1953).
30. Wang, M., Overland, J. E. & Bond, N. A. Climate projections for selected large marine ecosystems. *J. Mar. Syst.* **79**, 258–266 (2010).
31. Schlitzer, R. Electronic Atlas of WOCE Hydrographic and Tracer Data Now Available. *Eos Trans. AGU* **81**, 45 (2000).

**Acknowledgements** We thank J. Padman for assistance with faunal counts, K. Brewster for assistance with alkenone sample preparation and analysis, and A. Guiheneuf for preliminary alkenone measurements and faunal assemblage data. This work was supported by NSF grants AGS-0602395 (Project PALEOVAR, A.C.M.) and OCE-1204204 (A.C.M. and F.G.P.), and an NSF graduate research fellowship for S.K.P.; J.A.A. was supported by the USGS Climate and Land Use Change Research and Development Program and the Volcano Science Center.

**Author Contributions** S.K.P. and A.C.M. designed the study and wrote the paper. S.K.P., M.D.W., and F.G.P. contributed to alkenone palaeotemperature measurements and analysis. M.H.W. assisted with the chronology. J.A.A. provided insights on the trace metal and  $\delta^{15}\text{N}$  records. All authors contributed to interpretation of the data and provided comments on the manuscript.

**Author Information** The data can be found in the supplementary online materials and at the National Oceanic and Atmospheric Administration Paleoclimate Database. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.K.P. ([spraetorius@carnegiescience.edu](mailto:spraetorius@carnegiescience.edu)).



## METHODS

**Sediment cores.** Marine sediment core EW0408-85JC is located on the continental slope of the Gulf of Alaska (59° 33.32' N, 144° 9.21' W, 682 m). In the modern setting, this site lies near the upper margin of the OMZ, where oxygen concentrations are ~20  $\mu\text{mol kg}^{-1}$  (Fig. 1). The Gulf of Alaska margin experiences high seasonal productivity during the spring and late summer months, with the highest chlorophyll levels observed along the northern margin<sup>32</sup>, near the site of EW0408-85JC. The North Pacific drift feeds into the cyclonic Alaskan gyre and Alaskan Coastal Current (ACC), which drives downwelling along the margin<sup>33</sup>. Details and photographs of the sediment lithology for this core are previously published<sup>2</sup>.

Several sediment cores from various depths from the Northeast Pacific were also employed for the compilation of an oxygen isotope depth transect. Site EW0408-26JC/TC lies along the continental slope off the southeast Alaska margin (56° 96' N, 136° 43' W, 1,623 m), near the lower boundary of the OMZ (Fig. 1). Site EW0408-87JC is located under the Gulf of Alaska subpolar gyre, at a depth that lies underneath the modern OMZ (58° 77' N, 144° 50' W, 3,680 m).

**Age models.** The age model for core EW0408-85JC is based on 36 radiocarbon dates of mixed planktonic foraminifera<sup>20</sup> calendar corrected using the Bayesian radiocarbon chronology program BChron<sup>34</sup> with the Marine13 database<sup>35</sup>, assuming a marine reservoir correction of  $880 \pm 80$  yr. The age models for cores EW0408-26JC/TC consist of 10 radiocarbon dates on mixed planktonic foraminifera, calibrated with Calib 7.0 using a marine reservoir correction of  $735 \pm 50$  yr<sup>36</sup>. The age model for the trigger core of site EW0408-26 is poorly constrained due to low sedimentation rates, bioturbation, and carbonate dissolution in the upper sediments, therefore two tie points to the oxygen isotope stratigraphy of a nearby core with excellent age controls (EW0408-66JC) were used for the Holocene/YD boundary, and a modern age (0 yr BP) is assigned to the top of the core. The age model for core EW0408-87JC is based on 18 mixed planktonic radiocarbon dates calibrated with Calib 7.0 using a marine reservoir correction of  $850 \pm 100$  yr (Extended Data Table 1). One sample was excluded from the age model due to a small reversal. The age model for ODP Site 1019<sup>1</sup> (Extended Data Fig. 1) is updated with the age model from Lopes *et al.*<sup>17</sup>.

**Ventilation estimates.** Benthic radiocarbon ages were measured on mixed species of benthic foraminifera in the same samples as planktonic measurements to give an estimate of ventilation changes between the surface and deep waters at site EW0408-85JC<sup>20</sup>. Larger benthic-planktonic age differences reflect an increase in reservoir ages of subsurface waters, which may reflect either reduced ventilation rate, decreased preformed radiocarbon in the water mass, or mixing with an older water mass. Benthic ventilation ages were also evaluated using the projection age method<sup>37</sup>, which accounts for changes in the atmospheric  $^{14}\text{C}$  history based on the preformed radiocarbon content of surface waters before subduction, but does not account for subsurface mixing of multiple water masses.

**Oxygen isotopes.** Calculated  $\delta^{18}\text{O}$  of seawater ( $\delta^{18}\text{O}_{\text{sw}}$ ) combines planktonic  $\delta^{18}\text{O}$  of Nps or Gb<sup>2</sup> with alkenone palaeotemperatures (this paper) using a Gb calibration equation<sup>38</sup>, followed by correction for the global isotopic effect of changing ice volume<sup>39</sup>.

The depth transect of  $\delta^{18}\text{O}$  includes Nps  $\delta^{18}\text{O}$  data from core EW0408-85JC<sup>2</sup> as representative of near-surface conditions. It includes benthic  $\delta^{18}\text{O}$  representative of subsurface conditions in cores EW0408-85JC (682 m depth)<sup>2</sup> and EW0408-26JC/TC (1,623 m depth) from *Uvigerina peregrina* (Uvp), and in core EW0408-87JC (3,680 m depth) from *Cibicides wuellerstorfi* (+0.64 ‰). All were corrected for global ice volume<sup>39</sup>, but not for temperature.

**Alkenone palaeotemperature estimates.** Total lipids were extracted from ~5 g of freeze-dried sediment as per Walinsky *et al.*<sup>40</sup>. Linear, alkenone-containing fractions were isolated via urea adduction<sup>41</sup> and analysed using capillary gas chromatography with flame ionization detection. Analytical uncertainties in the quantification of C<sub>37</sub> ketone abundance (K37:2, K37:3 and K37:4) include both a mean potential 'baseline contaminant' component (positive) and an assumed 5% uncertainty in the integrated area (random). Minimum and maximum estimates of  $U_{37}^K$   $\{=(K37:2)/(K37:2 + K37:3)\}$ <sup>42</sup> were determined from the uncertainty in K37:2 and K37:3 concentrations. The corresponding uncertainties in temperature estimates are large in samples in the deepest part of the core (corresponding in time to 17.5–17.0 ka) due to extremely low K37 concentrations, most likely related high sedimentation rates. Core-top alkenone palaeotemperatures are biased towards summer, and thus are representative of temperatures experienced by the photosynthetic source of these biomarkers<sup>43</sup>.

Various studies have shown that different K37 compounds can be degraded selectively<sup>44</sup>. Consequently, the sea-surface water temperature proxy,  $U_{37}^K$ , can be diagenetically biased to some extent. Most lines of evidence indicate preferential degradation of the more unsaturated K37:3. As a result,  $U_{37}^K$  values, if altered, typically are shifted positively, thereby depicting apparently warmer values. The magnitude of the documented diagenetic warming bias appeared to be ~1 °C or less under the extreme conditions of alkenone degradation experienced in the

aerobic burn-down phenomena documented by turbidite records from the Madeira Abyssal Plain<sup>45</sup>.

Bacteria capable of both selective and non-selective aerobic degradation of alkenones have been studied in the laboratory<sup>46</sup>. In cases where selective alkenone degradation led to a positive shift in  $U_{37}^K$  values, epoxide derivatives were measured as intermediate products of the process. An empirical calibration of the apparent diagenetic 'warming' effect on  $U_{37}^K$  values caused as a function of these epoxide intermediates has now been defined<sup>45</sup>.

Alkenones and corresponding epoxide intermediates were measured in modern surface sediments collected by multi-core throughout our Southeast Alaska study area<sup>43</sup>. Interpretation of the results using the laboratory-defined empirical calibration suggested  $U_{37}^K$ -based SST estimates were biased too warm by as much as 1–2.5 °C<sup>45</sup>. The perceived warming effect shows a weak correlation ( $r = +0.26$ ) to the water depth at which each multi-core sample was collected, hinting that the magnitude of the warm bias is directly proportional to the availability of dissolved oxygen. Therefore, it is unlikely that estimated increases in SST during the hypoxic intervals in the Bolling–Allerød interstade (BA) and Holocene are related to changes in the sedimentary diagenetic environment. Observations of similar warming events during the BA and early Holocene from the California margin<sup>47</sup> (Extended Data Fig. 9) provide further support that these SST reconstructions reflect regional climate trends rather than local diagenetic imprints. Furthermore, the higher abundances of *Epistominella pacifica* relative to *Bolivina* and *Bulimina* species during cooling episodes in the SST reconstruction, such as the late glacial, Younger Dryas, and early to mid-Holocene (11–6 ka), suggest more oxygenated conditions during cool intervals. These trends are opposite of what would be expected if there was an influence of oxidation on the alkenone SST record, suggesting that a diagenetic warming effect, if present, would most likely act to dampen the observed magnitude of SST changes.

SST was also estimated via the  $U_{37}^K$  temperature index  $\{=(K37:2 - K37:4)/(K37:2 + K37:3 + K37:4)\}$  as calibrated by Prah *et al.*<sup>43</sup> (Extended Data Fig. 2). The  $U_{37}^K$  index has been suggested to be a more reliable proxy for SST at temperatures below 8 °C<sup>48,49</sup>. The  $U_{37}^K$  index results in SST estimates that are ~4 °C colder during the late glacial period and ~1 °C warmer during the Holocene period relative to the  $U_{37}^K$  SST estimates, reflecting the influence of changes in the concentration of tetraunsaturated K37:4, which tends to have higher abundances during glacial times relative to interglacial times (Prah *et al.*<sup>50</sup> and present study). However, the two SST estimates for the hypoxic intervals are virtually identical, providing further support that temperature estimates for these intervals are not influenced by preferential degradation of more unsaturated compounds. Additionally, the concentration of alkenones are highest during the hypoxic intervals (Extended Data Fig. 2), which not only increases the signal to noise ratio of our analyses (decreasing the associated SST errors), but is also consistent with excellent preservation of organic biomarkers during these events.

Average rates of SST change were calculated in a 400-yr window from the palaeotemperature record after interpolating on a 200-yr time step (Extended data Fig. 3).

**Benthic faunal abundances.** Benthic species abundances were counted from the >150  $\mu\text{m}$  size fraction with sample splits that ranged from 25–400 benthic specimens (Extended data Fig. 4). Individual specimens were classified into 12 genera and species categories, and the percent abundance for each species was calculated relative to the total number of benthic species counted. *Bulimina* and *Bolivina* are both elongate infaunal benthic genera that are considered indicators of low-oxygen conditions, and sometimes associated with high export productivity<sup>51–54</sup>. *Bulimina exilis* is indicative of the most severe hypoxic/anoxic conditions<sup>55</sup>. *Epistominella pacifica* is an epifaunal species that can tolerate intermediate to strong hypoxia and are often found in the upper and lower boundary zones of OMZs<sup>54</sup>. *Uvigerina peregrina* can tolerate intermediate to weak hypoxia, but is typically absent from the core of the OMZ. Therefore the relative abundances of these species reflect changes in bottom water oxygen concentrations, with high abundances of *Uvigerina peregrina* reflecting relatively well-oxygenated conditions, *Epistominella pacifica* reflecting intermediate hypoxia, *Bolivina* genera reflecting strong hypoxia, and *Bulimina exilis* reflecting strongly hypoxic to anoxic conditions.

**Trace metal data.** Trace metal concentration data and methods are previously published<sup>12,13</sup>. Total metal concentration data (Me<sub>T</sub>) of Mo and U was converted to excess (xs) values relative to lithogenic background, using Al concentration data in the core and the relationship:  $\text{Me}_{\text{xs}} = \text{Me}_{\text{T}} - (\text{Me}/\text{Al})_{\text{lithogenic}} \times \text{Al}_{\text{T}}$ , using average continental crust values of  $\text{Mo}/\text{Al} = 0.19 \times 10^{-4}$  and  $\text{U}/\text{Al} = 0.35 \times 10^{-4}$  (ref. 19). Both raw concentration data and excess concentrations of Mo and U show similar trends, with the greatest enrichments of U and Mo during the hypoxic intervals. **Organic data.** Biogenic silica and total organic carbon data and methods are previously published<sup>2,13</sup>. The  $\delta^{15}\text{N}$  data and methods are previously published, along with a detailed discussion of various influences that may contribute to the bulk  $\delta^{15}\text{N}$  signature<sup>13</sup>. A 'corrected' marine  $\delta^{15}\text{N}$  record was also calculated by

correcting for the  $\delta^{15}\text{N}$  component imparted from terrestrial organic matter<sup>13</sup> (Extended Data Fig. 5). This record shows similar overall trends to the raw  $\delta^{15}\text{N}$  record, with enriched  $\delta^{15}\text{N}$  during the BA and early Holocene hypoxic events. Elevated  $\delta^{15}\text{N}$  during the hypoxic intervals is consistent with an increase in nutrient utilization rate, which in this iron-limited setting would likely require an elevated iron source.

The  $\delta^{15}\text{N}$  data alone do not prove nutrient utilization; an alternate interpretation is that the high  $\delta^{15}\text{N}$  events during the deglacial transition are an advected signal from low latitudes, where a stronger oxygen minimum zone resulted in water column denitrification<sup>56</sup>. The viability of the undercurrent transport hypothesis is supported by modern tracer distributions that show California Undercurrent water detectable as far north as Alaska<sup>57</sup>. Addison *et al.*<sup>13</sup> discounted the undercurrent transport hypothesis as an explanation for deglacial increases in  $\delta^{15}\text{N}$ , simply because the fraction of undercurrent water reaching Alaska today is small (<15% of the subsurface water mass). Implausibly high variations in the tropical source waters (>12 ‰) would be required to explain the ~2‰  $\delta^{15}\text{N}$  changes in the Gulf of Alaska; such large changes have not been observed in the eastern tropical Pacific<sup>58</sup>. There is no dynamical mechanism to explain large increases in net pole-ward transport relative to mixing with ambient waters along the flow path that could yield such high-amplitude  $\delta^{15}\text{N}$  changes in the Gulf of Alaska.

Additional data argues in favour of a nutrient utilization mechanism to explain the high  $\delta^{15}\text{N}$  events in the Gulf of Alaska. Planktonic foraminiferal  $\delta^{13}\text{C}$  rises during the Bolling–Allerød and early Holocene warming events, sympathetic with high  $\delta^{15}\text{N}$  events (Extended Data Figs 5 and 6). If the  $\delta^{15}\text{N}$  were explained entirely by increased northward advection of high-nutrient, low-oxygen undercurrent waters from the tropics, the warm events should be accompanied by anomalously low  $\delta^{13}\text{C}$  of dissolved inorganic carbon. The same conflict appears for explanations of high productivity by upwelling of deep nutrient rich waters at these times<sup>5</sup>; upwelling of nutrient-rich waters without an increase in fractional nutrient utilization would yield low  $\delta^{13}\text{C}$  (Extended Data Fig. 6).

A regime change occurs in the early Holocene, from a deglacial interval in which  $\delta^{15}\text{N}$  is positively correlated with  $\delta^{13}\text{C}$  (17–9 ka,  $r^2 = 0.51$ ) to a Holocene system in which  $\delta^{15}\text{N}$  and  $\delta^{13}\text{C}$  are weakly negatively correlated (Extended Data Fig. 6). Consideration of changes in atmospheric  $\delta^{13}\text{C}$  (ref. 59) and air–sea equilibrium of the  $\delta^{13}\text{C}$  of carbonate ion as a function of temperature<sup>60</sup> strengthens this relationship (Extended Data Fig. 6). The co-occurrence of high  $\delta^{13}\text{C}$  with high  $\delta^{15}\text{N}$  during the deglacial interval points to nutrient utilization rate and carbon export as a key driver of the hypoxic events.

This view of nutrient utilization, likely related to removal of iron limitation during the deglacial interval is not inconsistent with the general view of advection of low-oxygen waters northward in the California Undercurrent. Indeed, northward advection of such low oxygen waters near the shelf-slope break would provide a potential reductive source of iron and phosphate<sup>24</sup> from sediments, and would be part of a self-sustaining feedback in the northeast Pacific in which initial hypoxia would be sustained and strengthened by iron-fuelled export productivity. Iron can be transported relatively long distances in the subsurface ocean in the colloidal (essentially non-sinking) fraction<sup>61</sup>. Such a mechanism is consistent with sea-level rise onto the continental shelves as a possible iron source<sup>1</sup>, and addresses concerns that isostatic rebound puts the local sea-level record in the northern Gulf of Alaska out of synch with the hypoxic events<sup>9</sup>.

**Leads/lags in proxy data.** The collection of multiple proxies within the same core allows for a precise examination of the timing of redox changes relative to changes in oceanographic conditions and export productivity (Extended Data Fig. 7). The transition to laminated sediments during the BA occurs with a sharp sedimentological boundary at 681 cm core depth<sup>2</sup>. The laminations are closely associated with high weight percentages of biogenic opal, consistent with high export of diatoms. The laminated intervals are also clearly defined in the X-ray computed tomography (CT) grey scan data, which largely tracks sediment density as a function of the biogenic to lithogenic fraction, with low values indicating times of high biogenic input<sup>2</sup>. The initial increase in SST precedes the increase in opal and the onset of laminations during the BA hypoxic event; temperatures >10 °C are associated with the interval of high diatom abundances. The increase in U and Mo occurs slightly before (2–5 cm) the increases in biogenic opal, TOC, and the decrease in CT grey scale. This may reflect a progression towards low oxygen conditions before the increase in export productivity. It is possible that such small depth offsets between the increase in trace metals and organic matter concentration may reflect preserved redox gradients in the sediment column, and thus not truly represent offsets in the time domain. However, all these proxies are emplaced within the bioturbated mixed layer and within a few cm of the seafloor during laminated intervals<sup>62</sup>, so depth offsets between proxies should be minor.

There are no discernible leads or lags between the increase in SST and export productivity for the Holocene hypoxic event. However, in this interval, the increase in U and Mo appear to slightly lag the increase in biogenic silica and the transition

to low-oxygen benthic fauna. The Holocene laminated interval is not as precisely defined as the onset of the BA laminations, partly due to weaker (slightly mottled) laminations, making the evaluation of depth offsets in the proxy data less reliable than for the BA sequence.

The switch from oxic to hypoxic/anoxic-tolerant benthic fauna occurs abruptly near the onset of laminations for both the BA and Holocene events. However, low-oxygen benthic species dominate the faunal assemblages well after the termination of laminations and the decrease in biogenic silica. Similar trends can be seen in the trace metal data. This concordance may indicate that low-oxygen conditions persisted even after the decline in export productivity and the cessation of laminations. This is consistent with a hysteresis-like response in the benthos, with an abrupt threshold transition to a hypoxic regime, followed by a more gradual return to an oxic regime with a diversity of benthic fauna<sup>63</sup>. Some upward mixing of older hypoxic fauna by bioturbation could have occurred when oxic conditions returned, however, this is unlikely to account for the high abundances (60–80%) of low-oxygen fauna that persisted in these intervals.

Based on this sequence of events, it appears most probable that sea surface warming lead to a reduction of dissolved oxygen in the subsurface through the combined effects of reduced oxygen solubility and enhanced thermal stratification. Benthic fauna assemblages from a deeper site in the Gulf of Alaska (near the lower boundary of the OMZ) suggest a gradual progression towards hypoxia starting at 16 ka, followed by an abrupt onset of laminations at the transition into the BA (Extended Data Fig. 8). The initial reduction of dissolved oxygen in the subsurface would have intensified the OMZ, leading to an expanded area of hypoxic shelf sediments (Fig. 3). A shoaling of the upper boundary of the OMZ to ~300 m during the BA hypoxic event has been documented in benthic faunal assemblages from the California Borderland basins<sup>64</sup>.

The supply of sedimentary iron is thought to be most efficient in a redox window where neither oxygen nor sulfide is present<sup>19</sup>, making such “new hypoxic zones” prime candidates for the release of bioavailable iron, especially in the shallower depths where iron can be more easily upwelled to the surface, as occurs in offshore regions of the Gulf of Alaska. Additional sources of iron include freshwater runoff charged with glacial rock flour. Such supplies of iron could have helped to fuel primary productivity, enhanced export productivity, and further depleted subsurface ocean concentrations, leading to a threshold-like feedback effect to amplify ocean deoxygenation.

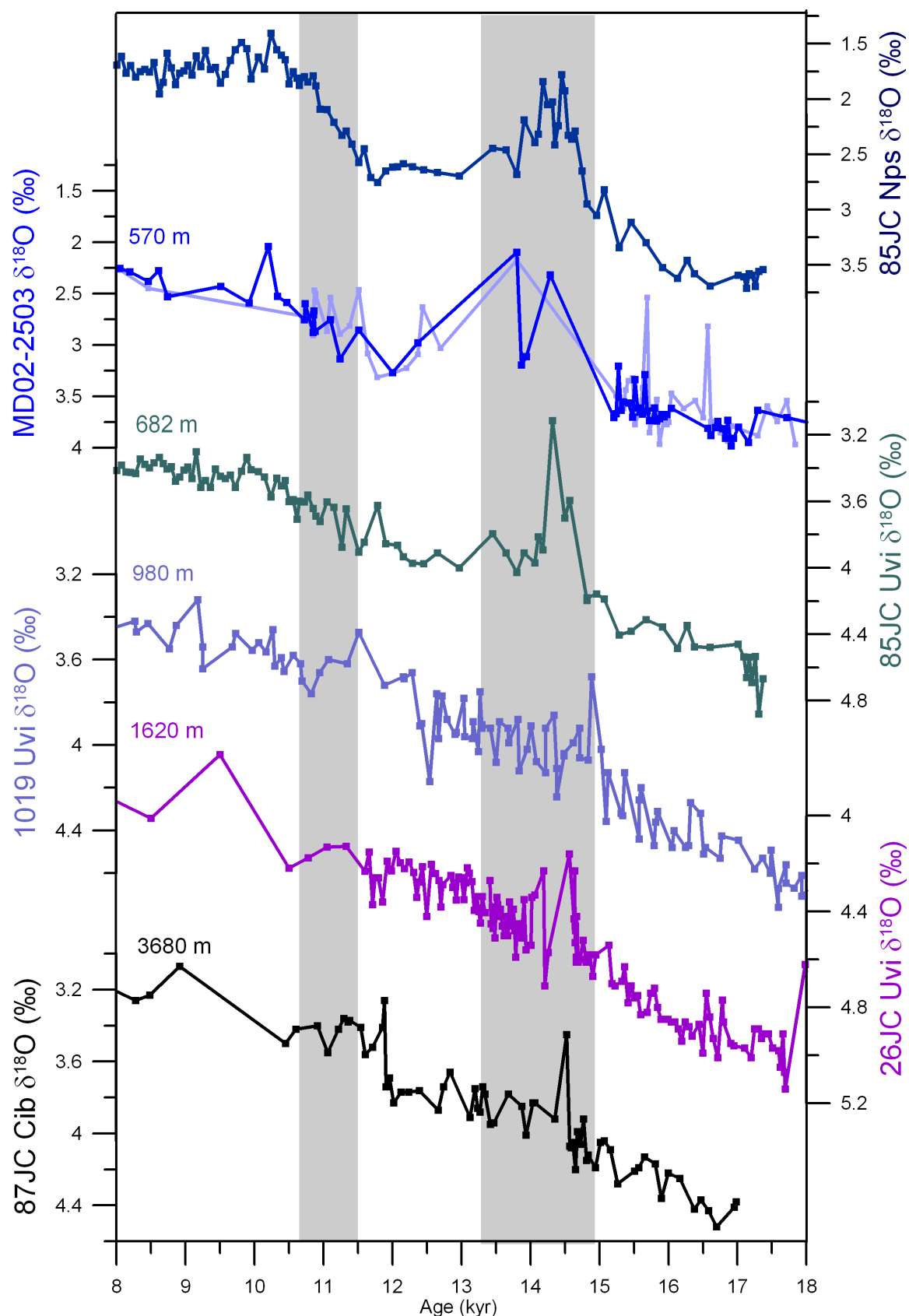
**Feedbacks on the carbon cycle.** The timing of the North Pacific hypoxic events approximately coincided with two intervals of abrupt increase in atmospheric  $\text{N}_2\text{O}$  (ref. 65) and a cessation in the rise of atmospheric  $\text{CO}_2$  (refs 66, 67) (Extended Data Fig. 9). The widespread expansion of hypoxic zones in the North Pacific could have led to denitrification, contributing to the two abrupt increases in atmospheric  $\text{N}_2\text{O}$ , while the enhanced export flux and burial of organic carbon may have helped to stabilize the rise in atmospheric  $\text{CO}_2$  (ref. 4). Thus, the temperature evolution of the North Pacific could play a prominent role in the regulation of multiple greenhouse gases.

Initial deglacial warming could promote out-gassing of deep-ocean respired carbon. Continuous and/or abrupt warming could have pushed large areas of the North Pacific across thresholds of hypoxia, in which the cycling of nutrients and carbon was fundamentally altered, contributing to expansive denitrification and the release of  $\text{N}_2\text{O}$  gases. However, the development of widespread hypoxia may ultimately act as a negative feedback on rising  $\text{CO}_2$  and global warming, with the release of nutrients from hypoxic sediments acting to stimulate surface productivity (in particular, diatoms with high efficiency for carbon export<sup>27</sup>) and the decrease in water column oxygen concentration helping to promote carbon burial (Fig. 3). Thus, thresholds of hypoxia in the North Pacific linked to ocean warming have the potential to switch this region from a source to sink of carbon.

32. Brickley, P. J. & Thomas, A. C. Satellite-measured seasonal and inter-annual chlorophyll variability in the Northeast Pacific and Coastal Gulf of Alaska. *Deep Sea Res. Part II Top. Stud. Oceanogr.* **51**, 229–245 (2004).
33. Stabeno, P. J. *et al.* Meteorology and oceanography of the Northern Gulf of Alaska. *Cont. Shelf Res.* **24**, 859–897 (2004).
34. Parnell, A. C. *et al.* A flexible approach to assessing synchronicity of past events using Bayesian reconstructions of sedimentation history. *Quat. Sci. Rev.* **27**, 1872–1885 (2008).
35. Reimer, P. J. *et al.* INTCAL13 and MARINE13 radiocarbon age calibration curves, 0–50,000 years cal BP. *Radiocarbon* **55**, 1869–1887 (2013).
36. Praetorius, S. K. & Mix, A. C. Synchronization of North Pacific and Greenland climates preceded abrupt deglacial warming. *Science* **345**, 444–448 (2014).
37. Adkins, J. F. & Boyle, E. A. Changing atmospheric  $\Delta^{14}\text{C}$  and the record of deep water paleoventilation ages. *Paleoceanography* **12**, 337–344 (1997).
38. Bemis, B. E., Spero, H. J., Bijma, J. & Lea, D. W. Reevaluation of the oxygen isotopic composition of planktonic foraminifera: Experimental results and revised paleotemperature equations. *Paleoceanography* **13**, 150–160 (1988).

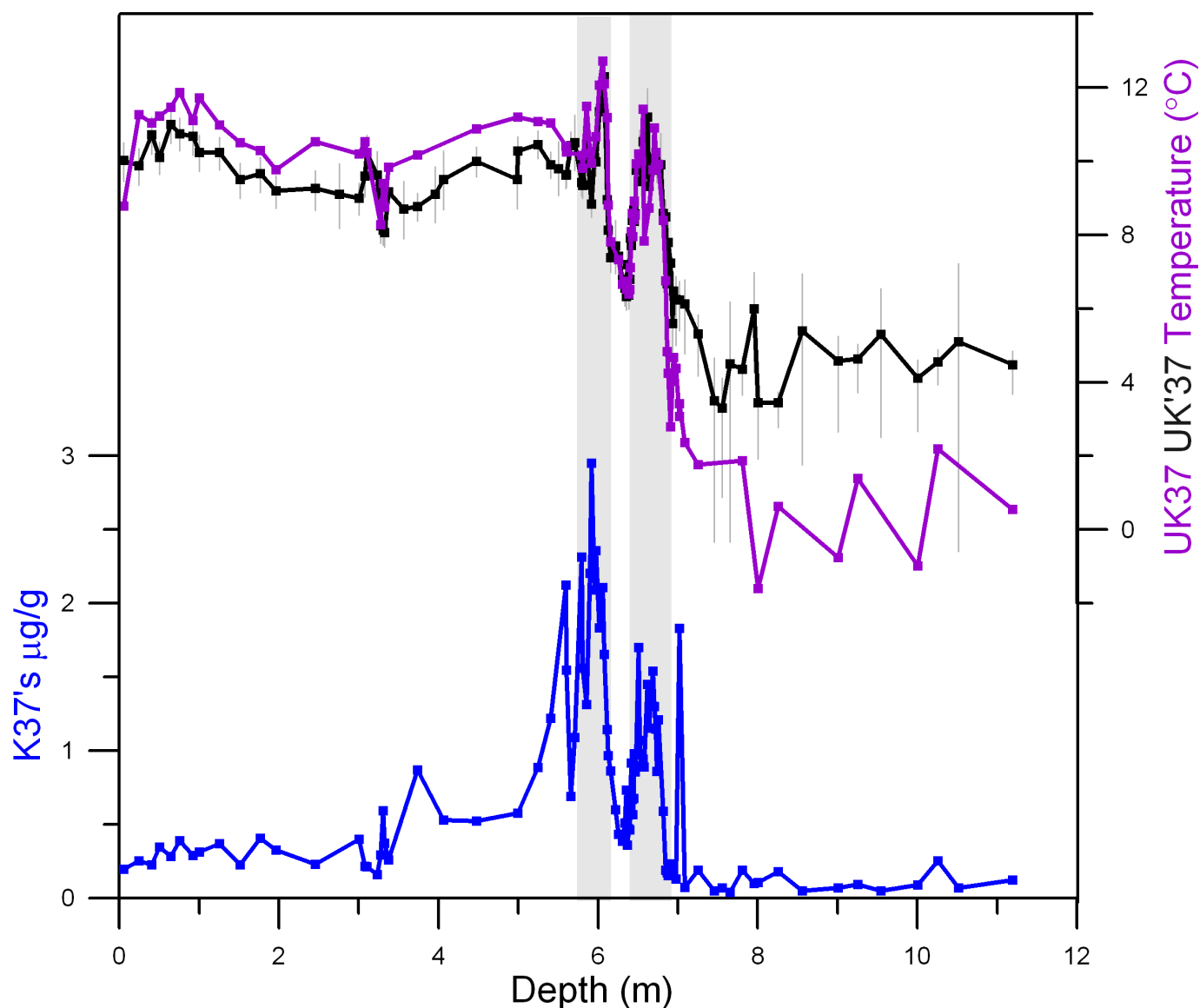
39. Waelbroeck, C. L. *et al.* Sea-level and deep water temperature changes derived from benthic foraminifera isotopic records. *Quat. Sci. Rev.* **21**, 295–305 (2002).
40. Walinsky, S. E. *et al.* Distribution and composition of organic matter in surface sediments of coastal Southeast Alaska. *Cont. Shelf Res.* **29**, 1565–1579 (2009).
41. Christie, W. W. & Han, X. *Lipid Analysis: Isolation, Separation, Identification and Structural Analysis of Lipids* 4th edn (Oily Press, 2003).
42. Prahl, F. G., Muehlhausen, L. A. & Zahnle, D. L. Further evaluation of long-chain alkenones as indicators of paleoceanographic conditions. *Geochim. Cosmochim. Acta* **52**, 2303–2310 (1988).
43. Prahl, F. G. *et al.* Systematic pattern in  $U_{37}^K$  – Temperature residuals for surface sediments from high latitude and other oceanographic settings. *Geochim. Cosmochim. Acta* **74**, 131–143 (2010).
44. Rontani, J. F., Volkman, J. K., Prahl, F. G. & Wakeham, S. G. Biotic and abiotic degradation of alkenones and implications on  $U_{37}^K$  paleoproxy application. *A review. Org. Geochem.* **59**, 95–113 (2013).
45. Prahl, F. G., Cowie, G. L., De Lange, G. J. & Sparrow, M. G. Selective organic matter preservation in “burn-down” turbidites on the Madeira Abyssal Pla in. *Paleoceanography* **18**, 1052 (2003).
46. Rontani, J. F. *et al.* Degradation of alkenones by aerobic heterotrophic bacteria: selective or not? *Org. Geochem.* **39**, 34–51 (2008).
47. Barron, J. A., Heusser, L., Herbert, T. & Lyle, M. High-resolution climatic evolution of coastal northern California during the past 16,000 years. *Paleoceanography* **18**, 1020 (2003).
48. Roselle-Mel , A. & Comes, P. Evidence for a warm Last Glacial Maximum in the Nordic seas or an example of shortcomings in  $U_{37}^K$  and  $U_{37}^K$  to estimate low seas surface temperature? *Paleoceanography* **14**, 770–776 (1999).
49. Bendle, J. & Roselle-Mel , A. Distributions of  $U_{37}^K$  and  $U_{37}^K$  in the surface waters and sediments of the Nordic Seas: Implications for paleoceanography. *Geochim. Geophys. Geosyst.* **5**, Q11013 (2004).
50. Prahl, F. G. *et al.* Assessment of sea-surface temperature at 42°N in the California Current over the last 30,000 years. *Paleoceanography* **10**, 763–773 (1995).
51. Bernhard, J. M. Characteristic assemblages and morphologies of benthic foraminifera from anoxia, organic-rich deposits: Jurassic through Holocene. *J. Foraminiferal Res.* **16**, 207–215 (1986).
52. Bernhard, J. & Reimers, C. Benthic foraminiferal population fluctuation related to anoxia: Santa Barbara Basin. *Biochemistry* **15**, 127–149 (1991).
53. Kaiho, K. Benthic foraminiferal dissolved-oxygen index and dissolved-oxygen levels in the modern ocean. *Geology* **22**, 719–722 (1994).
54. Jorissen, F. J., Fontanier, C. & Thomas, E. Paleocceanographical proxies based on deep-sea benthic foraminiferal assemblage characteristics. Proxies in Late Cenozoic Paleocceanography: Pt. 2: Biological tracers and biomarkers (eds Hillaire-Marcel C. & de Vernal, A.) 263–326 (Elsevier, 2007).
55. Hermelin, J. O. R. & Shimmield, G. B. The importance of the oxygen minimum zone and sediment geochemistry on the distribution of recent benthic foraminifera from the NW Indian ocean. *Mar. Geol.* **91**, 1–29 (1990).
56. Kienast, S. S., Calvert, S. E. & Pedersen, T. F. Nitrogen isotope and productivity variations along the northeast Pacific margin over the last 120 kyr: surface and subsurface paleocceanography. *Paleoceanography* **17**, 7–17 (2002).
57. Thomson, R. E. & Krassovski, M. V. Poleward reach of the California Undercurrent extension. *J. Geophys. Res.* **115**, C09027 (2010).
58. Robinson, R. S., Martinez, P., Pena, L. D. & Cacho, I. Nitrogen isotopic evidence for deglacial changes in nutrient supply in the eastern equatorial Pacific. *Paleoceanography* **24**, PA4213 (2009).
59. Schmitt, J. *et al.* Carbon isotope constraints on the deglacial CO<sub>2</sub> rise from ice cores. *Science* **336**, 711–714 (2012).
60. Zhang, J., Quay, P. D. & Wilbur, D. O. Carbon isotope fractionation during gas water exchange and dissolution of CO<sub>2</sub>. *Geochim. Cosmochim. Acta* **59**, 107–114 (1995).
61. Fitzsimmons, J. N., Boyle, E. A. & Jenkins, W. J. Distal transport of dissolved hydrothermal iron in the deep South Pacific Ocean. *Proc. Natl Acad. Sci. USA* **111**, 16654–16661 (2014).
62. Chang, A. S., Pichevin, L., Pedersen, T. F., Gray, V. & Ganeshram, R. New insights into productivity and redox-controlled trace element (Ag, Cd, Re, and Mo) accumulation in a 55 kyr long sediment record from Guaymas Basin, Gulf of California. *Paleoceanography* **30**, 77–94 (2015).
63. Conley, D. J., Carstensen, J., Vaquer-Sunyer, R. & Duarte, C. M. Ecosystem thresholds with hypoxia. *Hydrobiologia* **629**, 21–29 (2009).
64. Moffitt, S. E., Hill, T. M., Ohkushi, K., Kennett, J. P. & Behl, R. Vertical oxygen minimum zone oscillations since 20 ka in Santa Barbara Basin: A benthic foraminiferal community perspective. *Paleoceanography* **29**, 44–57 (2014).
65. Schilt, A. *et al.* Atmospheric nitrous oxide during the last 140,000 years. *Earth Planet. Sci. Lett.* **300**, 33–43 (2010).
66. Monnin, E. *et al.* Atmospheric CO<sub>2</sub> concentrations over the last glacial termination. *Science* **291**, 112–114 (2001).
67. Marcott, S. A. *et al.* Centennial-scale changes in the global carbon cycle during the last deglaciation. *Nature* **514**, 616–619 (2014).
68. Hill, T. M. *et al.* Pre-B lling warming in Santa Barbara Basin, California: surface and intermediate water records of early deglacial warmth. *Quat. Sci. Rev.* **25**, 2835–2845 (2006).





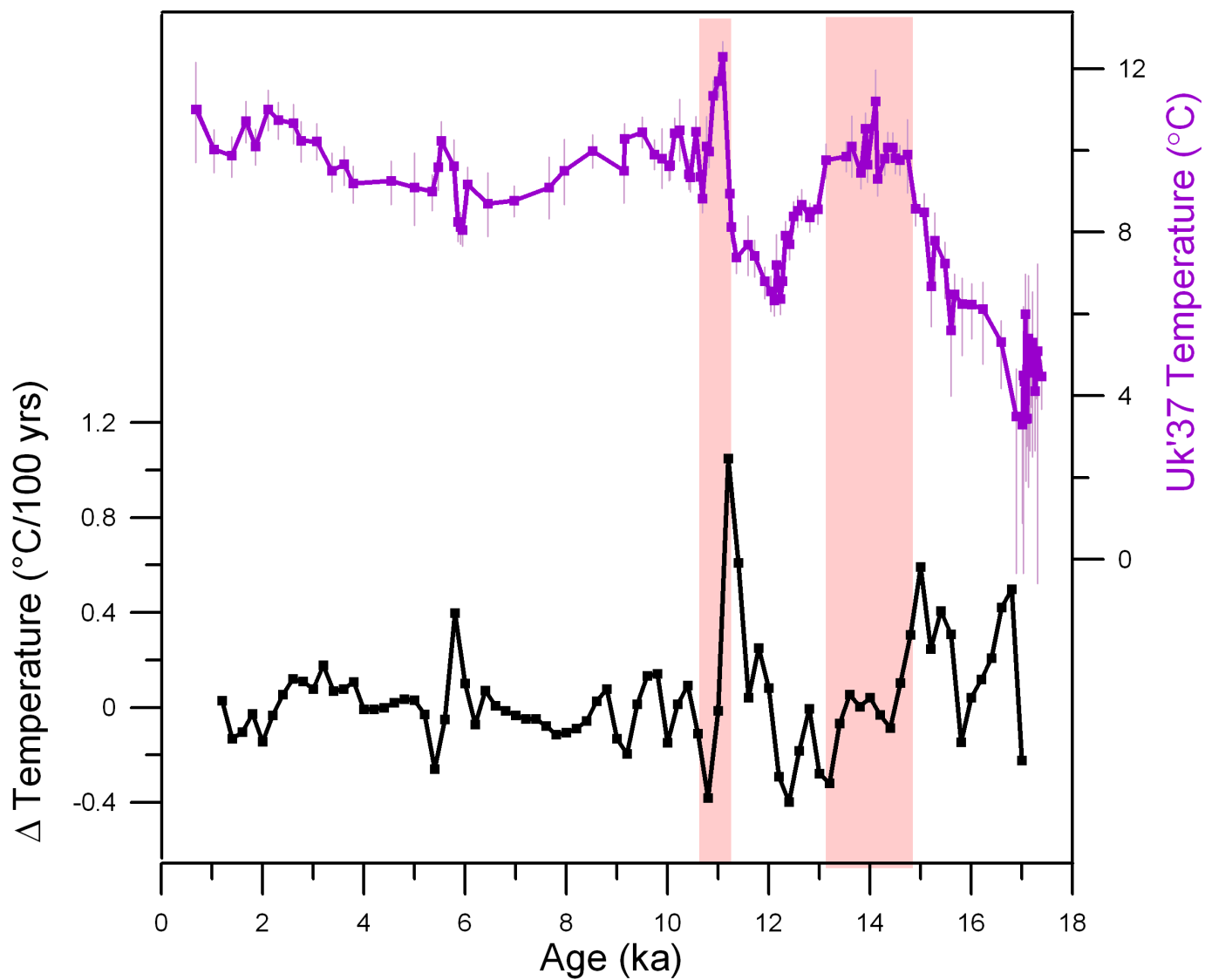
**Extended Data Figure 1 | Depth transect of oxygen isotopes from the Northeast Pacific.** Planktonic oxygen isotopes (Nps) from core EW0408-85JC (dark blue)<sup>2</sup>, benthic oxygen isotopes from core MD02-2503 in the Santa Barbara basin (*Uvigerina peregrina*; light blue, *Bolivina argentea*; bright blue)<sup>68</sup>, benthic *Uvigerina peregrina* oxygen isotopes

from core EW0408-85JC (green)<sup>2</sup>, ODP Site 1019 ((41° 68' N, 124° 93' W, 978 m; light blue)<sup>1</sup>, cores EW0408-26JC/TC, and core EW0408-87JC (*Cibicides*). Data from the Gulf of Alaska cores (EW0408) are used to make the depth-time map shown in Fig. 4.



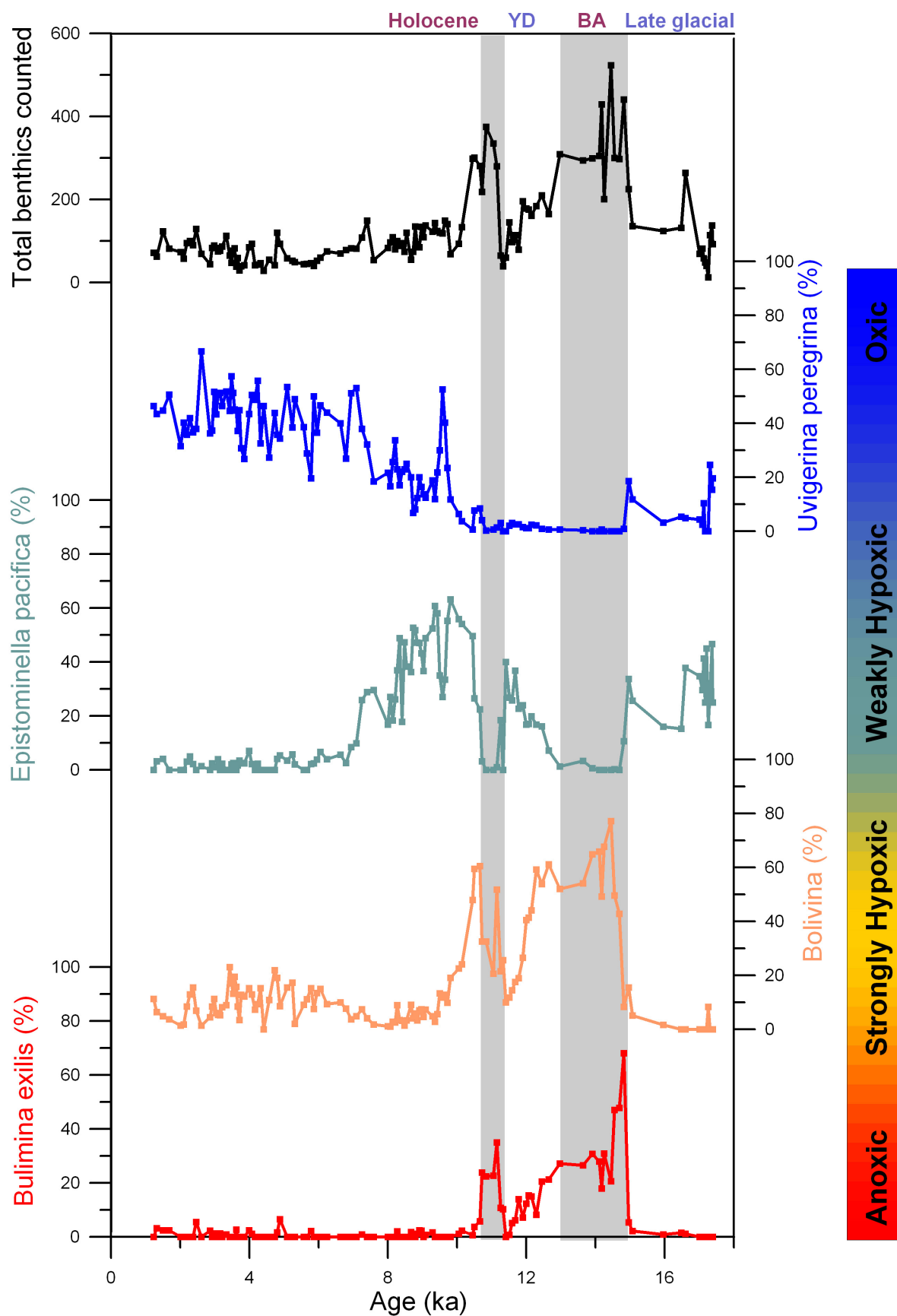
**Extended Data Figure 2 | Comparison of SST estimates based on the  $U_{37}^K$  (purple) and  $U_{37}'^K$  (black) indices.** Temperatures based on the  $U_{37}^K$  index show a larger glacial-interglacial change, with colder SSTs during the late glacial period and warmer SST for the Holocene. Temperature estimates for the deglacial period, including the two hypoxic intervals and the Younger

Dryas, are virtually identical between methods, giving confidence that diagenetic effects are not influencing the alkenone ratios during the hypoxic warm events. Alkenone concentrations are high during the two hypoxic events (blue), consistent with other proxy evidence for high productivity and/or excellent preservation of organic matter during these events.



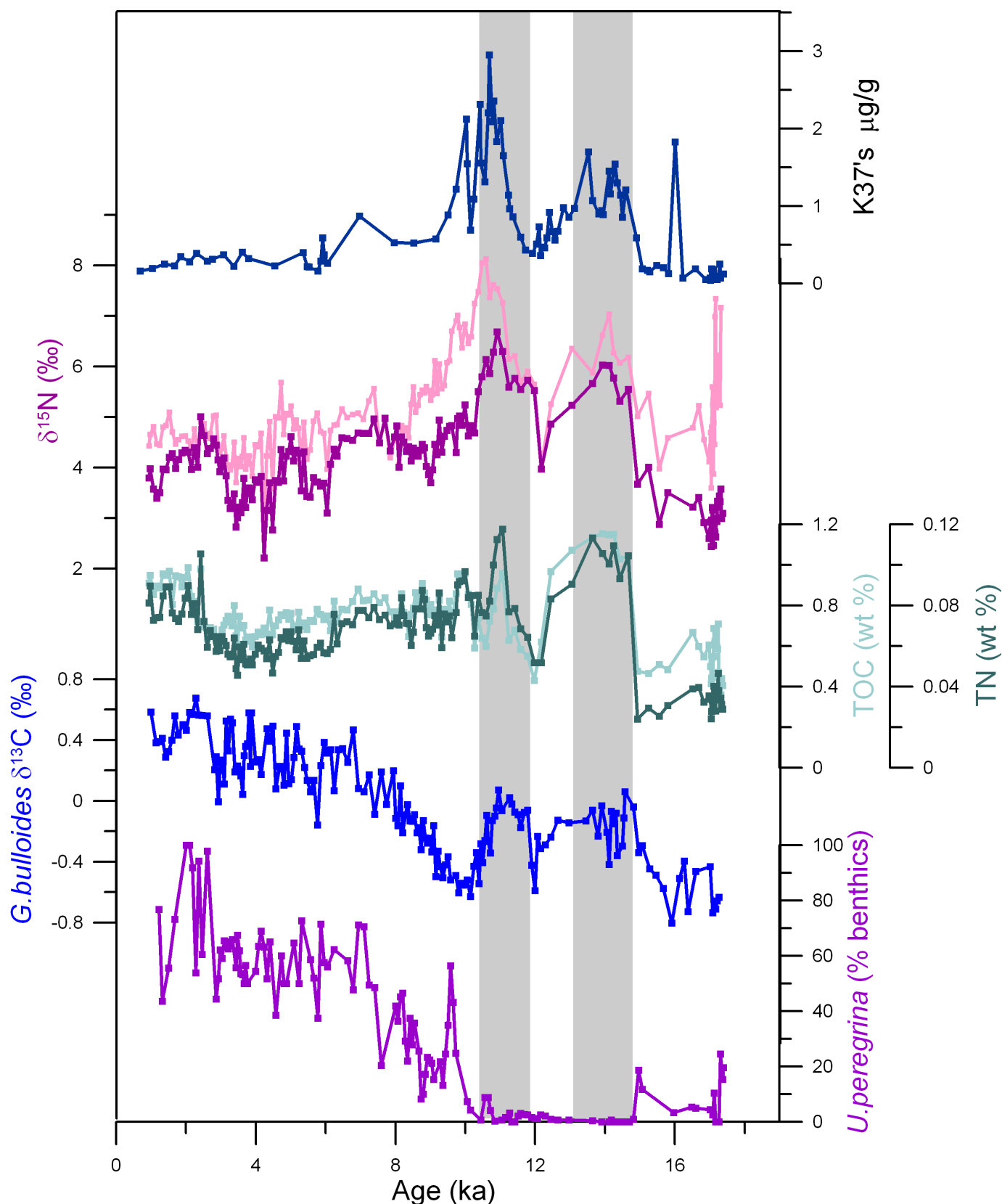
**Extended Data Figure 3 | Rate of SST change in the Gulf of Alaska.** The alkenone palaeotemperature record (purple) was interpolated on a 200-yr time step and the average rate of temperature change ( $^{\circ}\text{C}$  per century) was calculated over a 400-yr window (black).





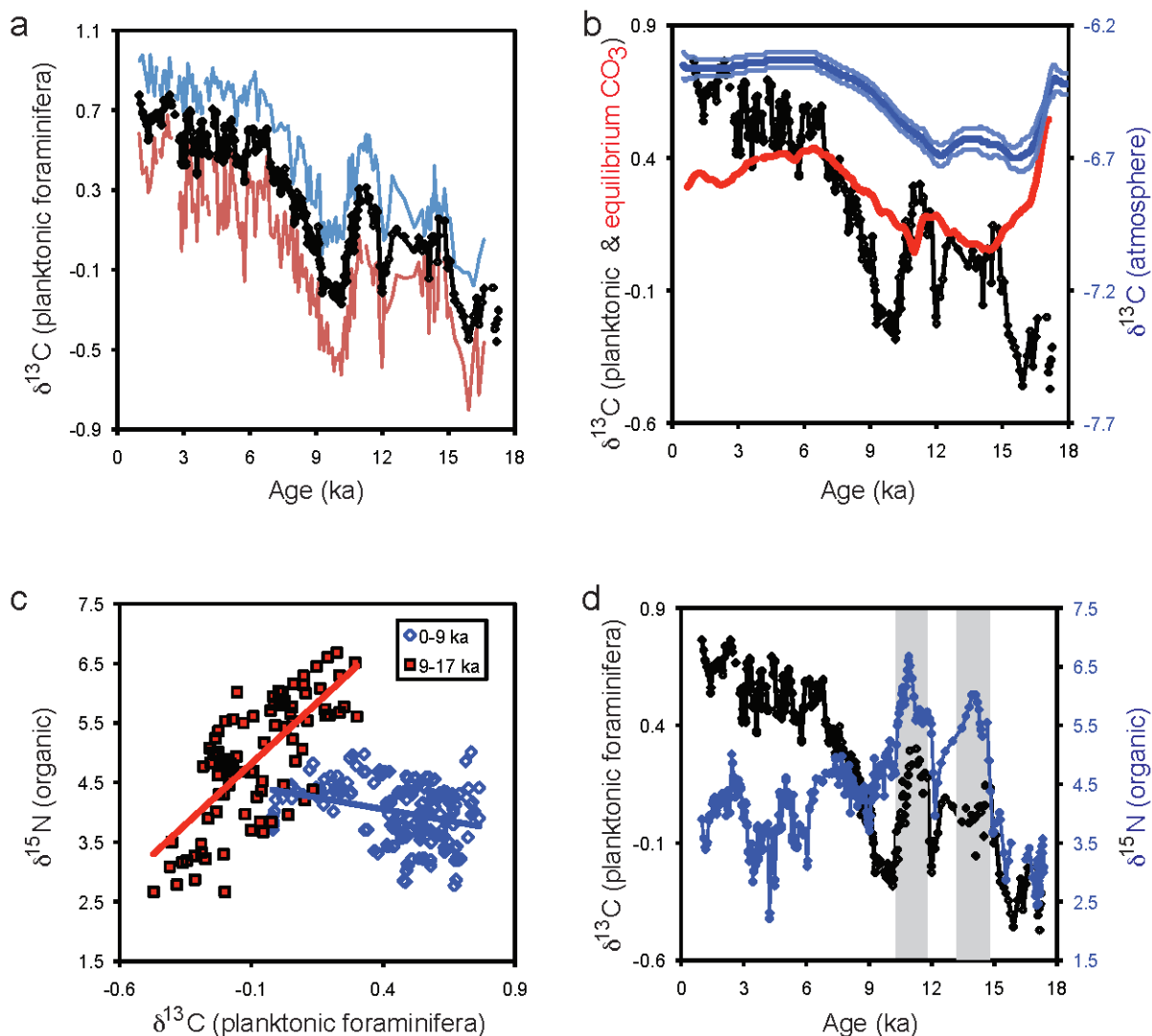
**Extended Data Figure 4 | Relative abundances of benthic species and genera in core EW0408-85JC.** *Bulimina exilis* is the most tolerant of low-oxygen conditions, and is often associated with near-anoxic bottom waters. *Bulimina* and *Bolivina* genera are typically found in strongly hypoxic

waters, whereas *Epistominella pacifica* is associated with intermediate hypoxia, and *Uvigerina peregrina* is associated with more well-oxygenated conditions. Grey shaded bars represent the two laminated intervals, which are almost exclusively comprised of *Bolivina* and *Bulimina* genera.



**Extended Data Figure 5 | Records of surface and export productivity from EW0408-85JC.** The sedimentary  $\delta^{15}\text{N}$  record (violet) and  $\delta^{15}\text{N}$  corrected for terrestrial organic matter (light violet)<sup>13</sup> show elevated values during the two hypoxic intervals (grey bars), which also coincide with enhanced organic matter deposition, including total organic carbon (light green)<sup>13</sup> and an increase in the alkenone K37 abundance (dark blue).

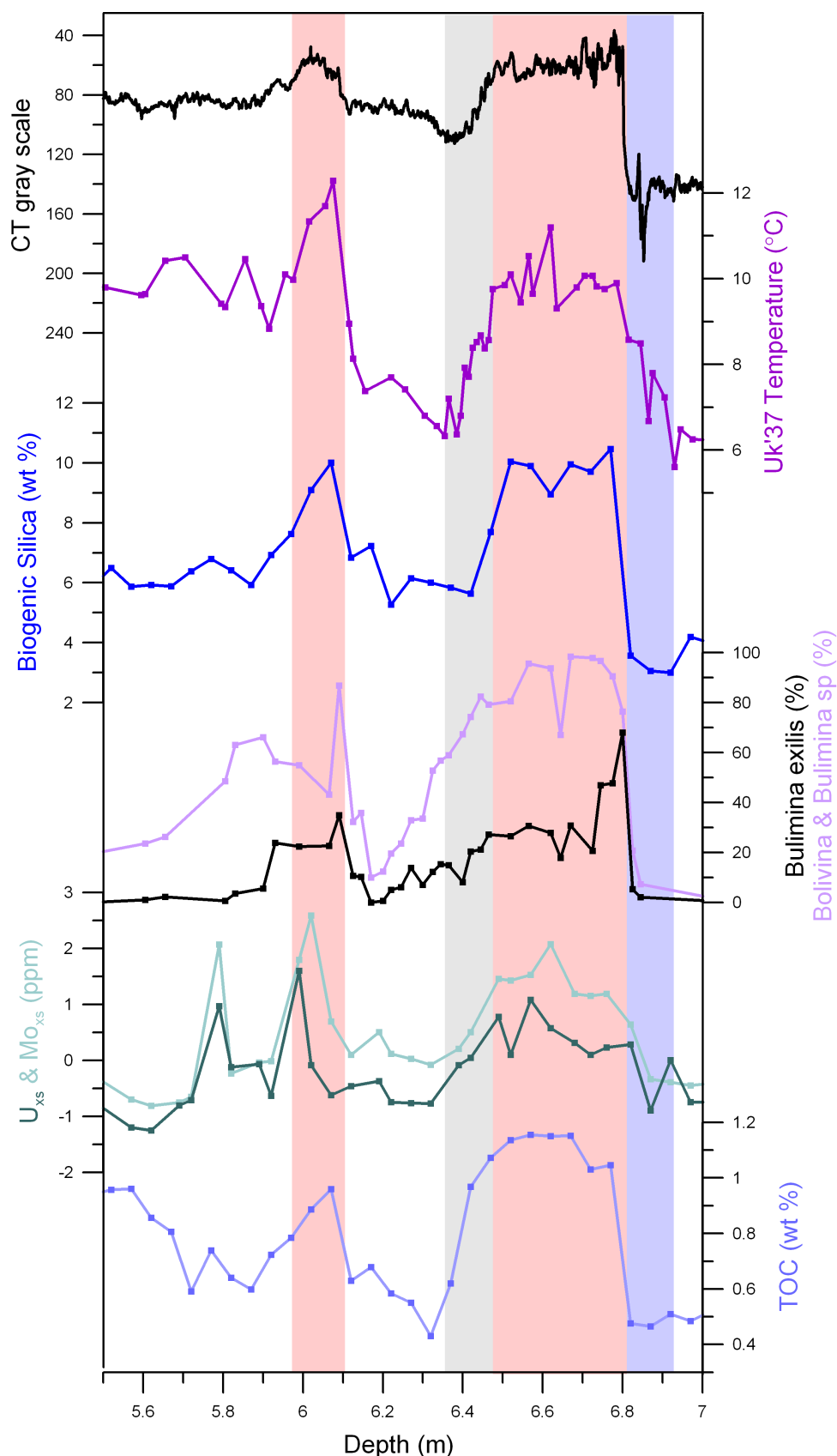
High planktonic  $\delta^{13}\text{C}$  values are observed during these intervals (bright blue)<sup>2</sup>, consistent with an increase in surface productivity rather than upwelling of deep waters exported from low-latitudes. The progressive increase in planktonic  $\delta^{13}\text{C}$  through the Holocene is accompanied by an increase in the relative abundance of *U. peregrina* (purple), likely indicating a better ventilated water column in the Holocene.



**Extended Data Figure 6 | Comparison of planktonic  $\delta^{13}\text{C}$  with sedimentary  $\delta^{15}\text{N}$  in core EW0408-85JC.** **a**,  $\delta^{13}\text{C}$  of planktonic foraminifera, Nps (blue), Gb (red), average of the two species (black). **b**, Comparison of the average planktonic foraminiferal  $\delta^{13}\text{C}$  (black) with changes in  $\delta^{13}\text{C}$  of atmospheric  $\text{CO}_2$  (blue)<sup>57</sup> and estimated surface ocean  $\delta^{13}\text{C}$  of  $\text{CO}_3^{2-}$  (red, calculated from the smooth atmospheric values using the temperature relationship of Zhang *et al.*<sup>60</sup>). **c**, Relationship between  $\delta^{15}\text{N}$  in organic matter and  $\delta^{13}\text{C}$  in planktonic foraminifera (average of Gb and Nps). The  $\delta^{15}\text{N}$  and  $\delta^{13}\text{C}$  measurements were in most cases made in adjacent samples, typically separated by 1 cm. To prevent directional bias in the scatter plot, the two variables were first interpolated linearly onto

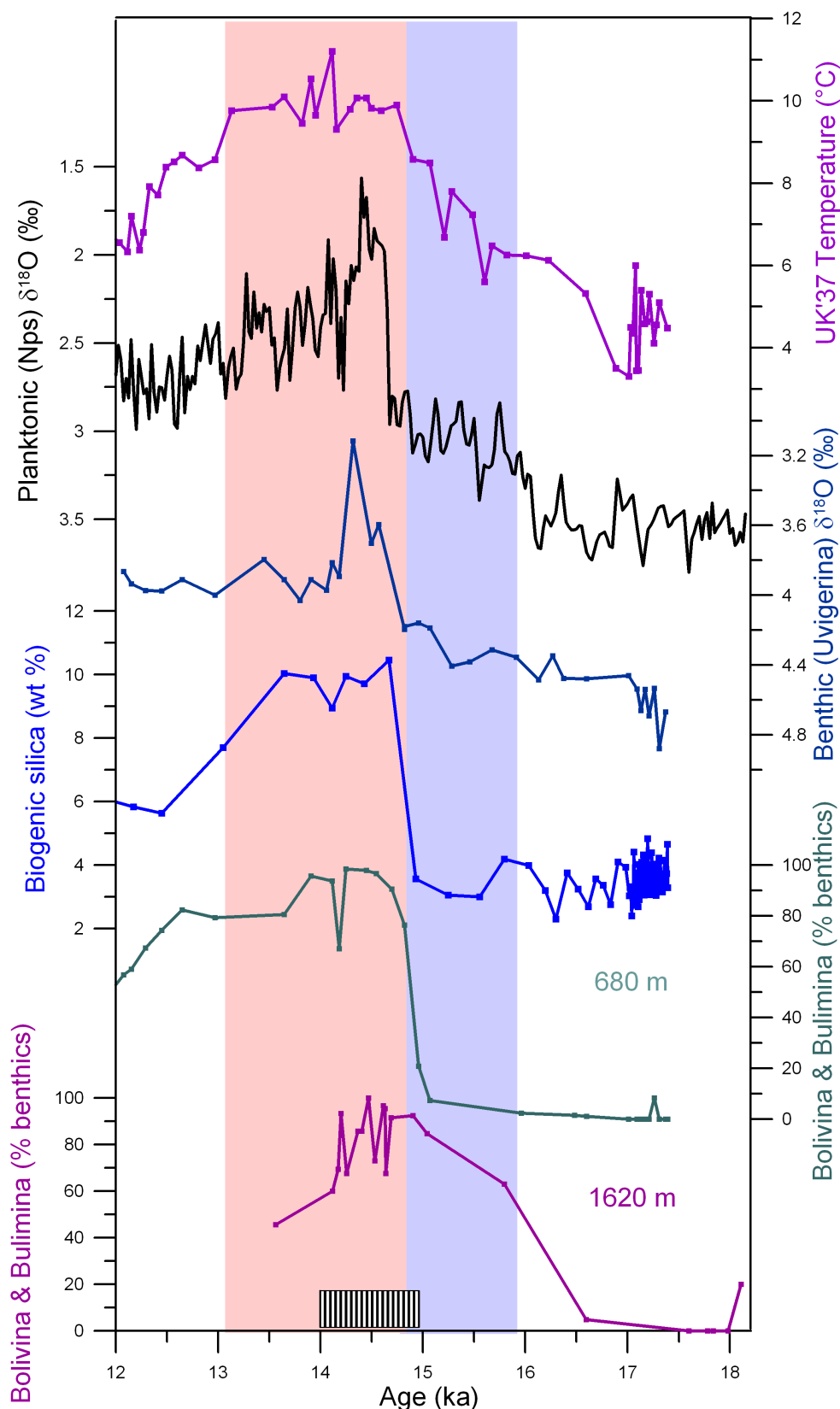
the depth of the other variable. In no cases were interpolations allowed over an interval  $>5$  cm or 200 years. The positive correlation between organic  $\delta^{15}\text{N}$  and planktonic foraminiferal  $\delta^{13}\text{C}$  during the deglacial interval (17–9 ka, red points,  $r^2 = 0.51$ ) supports an interpretation of increased nutrient utilization and carbon export from near-surface waters associated with the high  $\delta^{15}\text{N}$  events. In contrast, within Holocene time (9–0 ka) the relationship between  $\delta^{15}\text{N}$  and  $\delta^{13}\text{C}$  reverses, suggesting no systematic variations in nutrient utilization. **d**, Time series of organic  $\delta^{15}\text{N}$  (organic) and  $\delta^{13}\text{C}$  (average of the planktonic foraminifera Gb and Nps) as a time series; these data form the comparison in **c**.





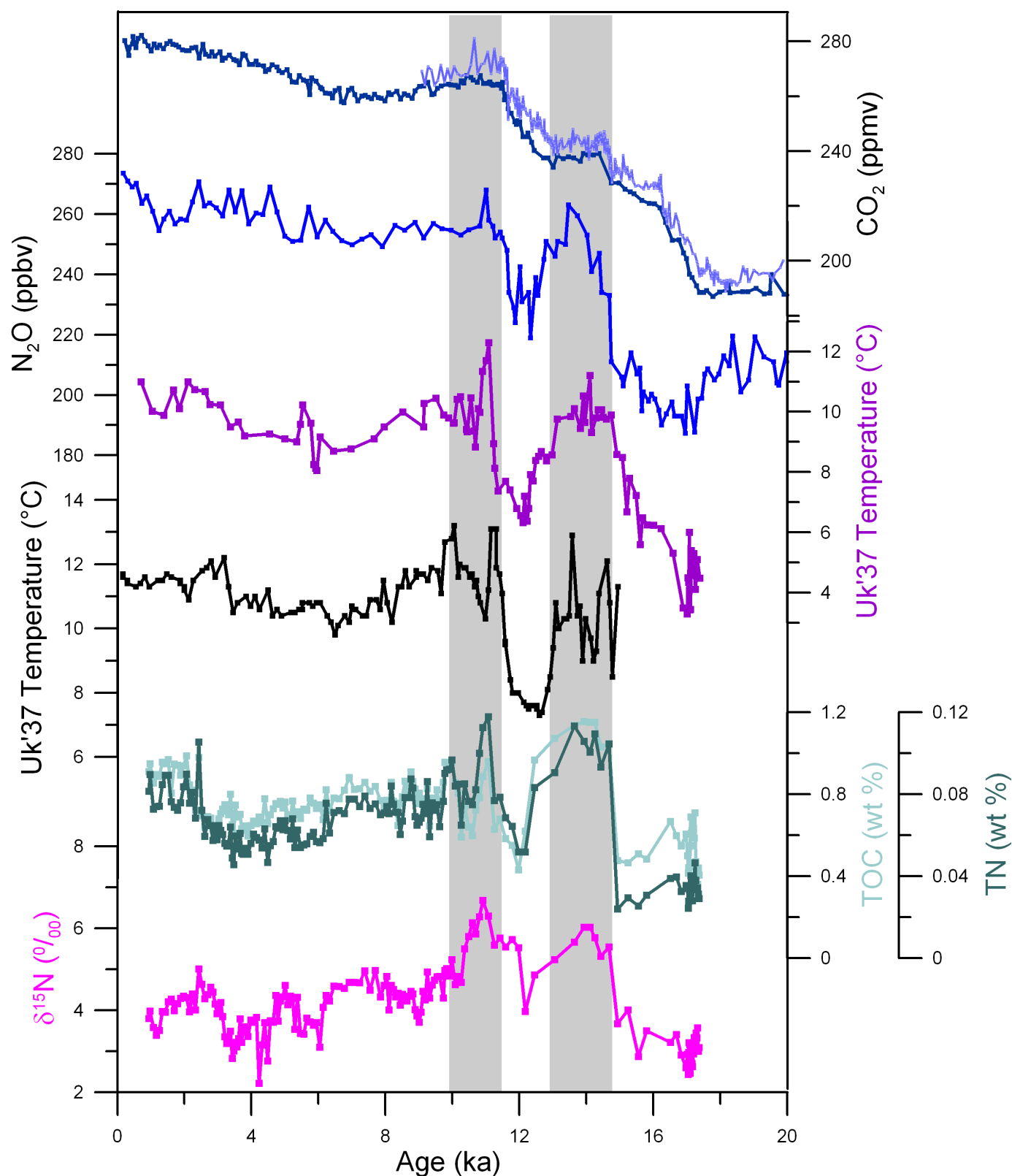
**Extended Data Figure 7 | Expanded view of proxy data (plotted as depth in core) for EW0408-85JC during the two hypoxic intervals.** CT grey scale (black) reflects changes in the biogenic:lithogenic fraction of sediment, with low values indicating times of high biogenic input (primarily diatoms)<sup>2</sup>. The laminated intervals (pink shading) coincide with high diatom abundance and SSTs near or exceeding 10 °C, whereas

evidence for low-oxygen conditions appears to extend both before (blue shading) and after (grey shading) the BA laminated zone based on trace metal concentrations<sup>12,13</sup>, benthic faunal assemblages, and preservation of TOC<sup>13</sup>, coinciding with the initial increase in SST and decrease in benthic  $\delta^{18}\text{O}$  (dark blue)<sup>2</sup>.



**Extended Data Figure 8 | Surface climate proxies compared with changes in benthic  $\delta^{18}\text{O}$  and fauna from different depth sites in the Gulf of Alaska.** The alkenone palaeotemperature record from core EW0408-85JC (purple), a composite record of planktonic  $\delta^{18}\text{O}$  from cores EW0408-26JC and EW0408-66JC (black)<sup>36</sup>, benthic  $\delta^{18}\text{O}$  (dark blue) and biogenic silica (bright blue) from core EW0408-85JC<sup>2</sup>, the combined abundance of low-oxygen tolerant *Bolivina* and *Bulimina* species from cores EW0408-85JC (682 m; green) and EW0408-26JC (1,620 m; violet). An increase in

low-oxygen benthic fauna is apparent in the deeper site (EW0408-26JC) commencing at 16 ka, which coincides with the pre-Bølling warming in the SST record and an increase in the planktonic and benthic  $\delta^{18}\text{O}$  records. This initial decrease (blue shading) in sedimentary oxygen content at the base of the OMZ clearly precedes the large increase in biogenic silica and the shift to hypoxic conditions in core EW0408-85JC near the onset of sedimentary laminations (pink shading). Sediment laminations in core EW0408-26JC occur from 15–14 ka (shaded bar on x-axis).



**Extended Data Figure 9 | Northeast Pacific SSTs, productivity indices, and atmospheric greenhouse gases.** Data from top: CO<sub>2</sub> record from EDC (dark blue)<sup>66</sup> and WAIS (light blue)<sup>67</sup>, a record of N<sub>2</sub>O from TALOS Dome (bright blue)<sup>65</sup>, the Gulf of Alaska U<sub>37</sub>K' SST record (purple), a U<sub>37</sub>K' SST record from the California margin (black)<sup>47</sup>, records of total organic carbon (TOC:

light green), total nitrogen (TN: dark green), and δ<sup>15</sup>N records on bulk organic matter from core EW0408-85JC<sup>13</sup>. Grey shaded bars indicate the two intervals in which the deglacial rise in CO<sub>2</sub> plateaus/reverses, which generally correspond to the episodes of widespread North Pacific hypoxia, high SSTs, enhanced nitrate utilization, and increased export productivity.



Extended Data Table 1 | Radiocarbon age controls for core EW0408-87JC

Depth midpoint (cm)	Planktonic $^{14}\text{C}$ age (yr)	$^{14}\text{C}$ +/- (yr)	Marine13 age (yr)	2 sigma error (yr)
39.0	1640	15	760	110
208.5	8520	20	8530	140
245.0	10715	20	11330	160
250.0	10740	40	11380	240
255.0	10975	30	11700	300
267.0	11090	25	11990	280
281.0	11695	30	12740	130
301.0	12460	40	13470	160
329.0	13170	40	14370	320
341.0	13330	45	14650	400
381.0	13830	30	15510	230
420.5	14290	40	16160	230
460.0	14560	70	16570	310
492.0	14840	70	16950	330
540.0	14930	60	17120	300
920.0	15060	45	17300	220
1002.0	15020	50	17250*	240
1449.0	15445	40	17770	190

Radiocarbon measurements were made on mixed planktonic foraminifera, sinistral *Neogloboquadrina pachyderma* (Nps) and *Globigerina bulloides* (Gb) picked from the >150  $\mu\text{m}$  size fraction. Radiocarbon samples were analysed at the Keck AMS facility at U.C. Irvine. Radiocarbon dates were calibrated with Calib 7.0 using the Marine13 calibration curve and a marine reservoir correction of  $850 \pm 100$  yr. The 2- $\sigma$  midpoint age, rounded to the nearest decade, is used for the age model. One sample was excluded from the age model due to a minor age reversal (denoted with an asterisk), although this date is within uncertainty of the adjacent shallower radiocarbon date, indicating very high sedimentation rates in this section of the core.

# The effects of life history and sexual selection on male and female plumage colouration

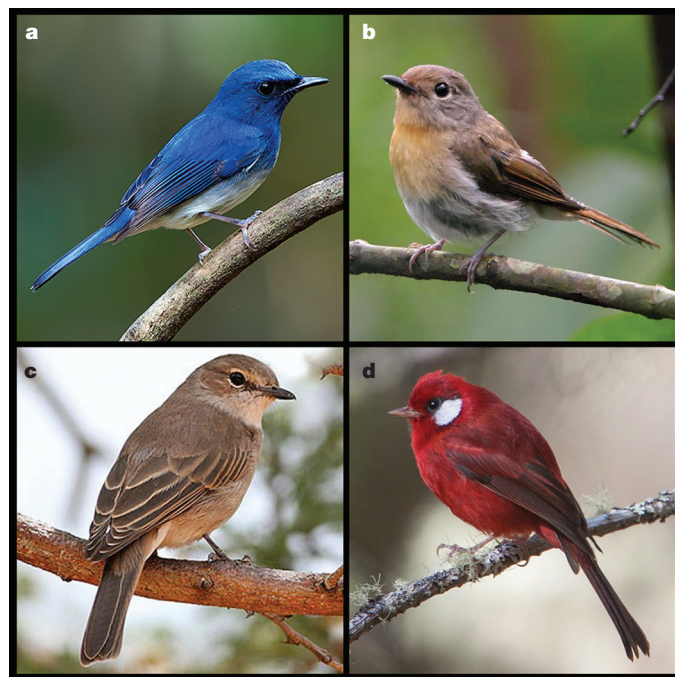
James Dale<sup>1</sup>, Cody J. Dey<sup>2</sup>, Kaspar Delhey<sup>3,4</sup>, Bart Kempenaers<sup>5</sup> & Mihai Valcu<sup>5</sup>

Classical sexual selection theory<sup>1–4</sup> provides a well-supported conceptual framework for understanding the evolution and signalling function of male ornaments. It predicts that males obtain greater fitness benefits than females through multiple mating because sperm are cheaper to produce than eggs. Sexual selection should therefore lead to the evolution of male-biased secondary sexual characters. However, females of many species are also highly ornamented<sup>5–7</sup>. The view that this is due to a correlated genetic response to selection on males<sup>1,8</sup> was widely accepted as an explanation for female ornamentation for over 100 years<sup>5</sup> and current theoretical<sup>9,10</sup> and empirical<sup>11–13</sup> evidence suggests that genetic constraints can limit sex-specific trait evolution. Alternatively, female ornamentation can be the outcome of direct selection for signalling needs<sup>7,14</sup>. Since few studies have explored interspecific patterns of both male and female elaboration, our understanding of the evolution of animal ornamentation remains incomplete, especially over broad taxonomic scales. Here we use a new method to quantify plumage colour of all ~6,000 species of passerine birds to determine the main evolutionary drivers of ornamental colouration in both sexes. We found that conspecific male and female colour elaboration are strongly correlated, suggesting that evolutionary changes in one sex are constrained by changes in the other sex. Both sexes are more ornamented in larger species and in species living in tropical environments. Ornamentation in females (but not males) is increased in cooperative breeders—species in which female–female competition for reproductive opportunities and other resources related to breeding may be high<sup>6</sup>. Finally, strong sexual selection on males has antagonistic effects, causing an increase in male colouration but a considerably more pronounced reduction in female ornamentation. Our results indicate that although there may be genetic constraints to sexually independent colour evolution, both female and male ornamentation are strongly and often differentially related to morphological, social and life-history variables.

The extraordinary interspecific variation in bird colouration has provided model studies on animal ornamentation (Fig. 1). The many striking cases of sexual dichromatism in birds illustrate the power of sexual selection because such species often have highly polygynous mating systems<sup>1,4</sup>. However, other factors besides sexual selection can influence colour elaboration<sup>15,16</sup> and there is growing evidence that many female ornaments are adaptive and subject to direct selection<sup>5,7</sup>. Females often compete for ecological resources, and may use ornamental traits to mediate competitive interactions. Since male ornaments are also used during competition for non-sexual resources (for example, food, territories), ornament evolution in both sexes may be better understood through the concept of social selection<sup>14</sup>, which encompasses both traditional sexual selection, and selection on non-sexual interactions<sup>7,17</sup>. Under this framework, male and female ornamentation should correlate with both sexual selection and life-history traits that

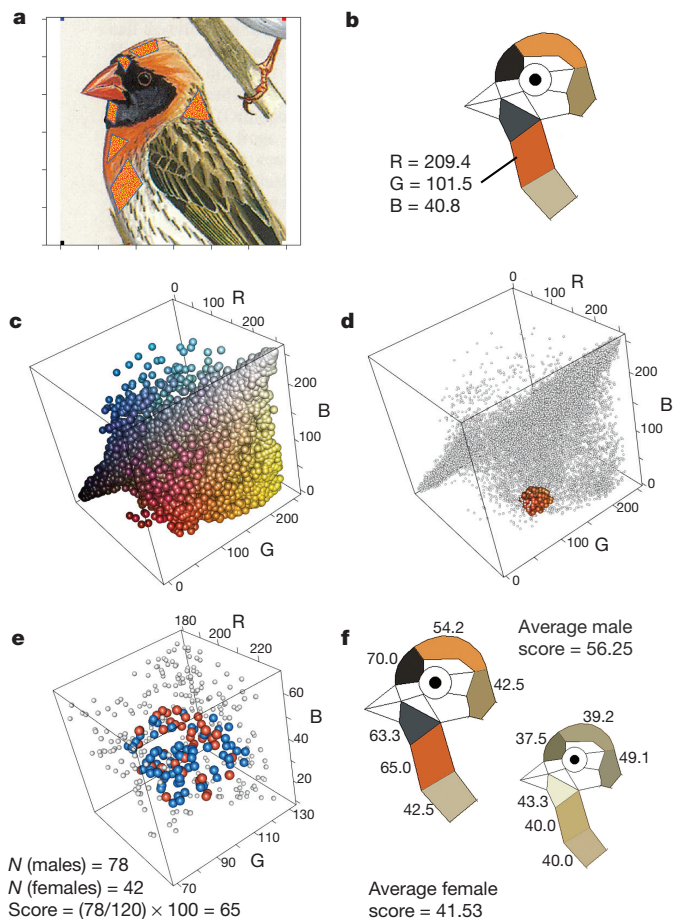
potentially influence the level of competition for resources. To date, the life-history traits associated with ornamentation in both males and females remain poorly understood.

A fundamental challenge in resolving these issues is the quantification of colour ornamentation in a way that allows meaningful interspecific comparisons. We developed a method to quantify colour elaboration by determining how ‘male-like’ a focal plumage is (Fig. 2). This approach, which can be used with any colour quantification technique (see Methods and Extended Data Fig. 1), has the important property of quantifying diverse colours using a single metric (that is, how male-like it is). Thus, birds with dramatically different appearances and of different sex can have similar scores (for example, Fig. 1a, d).



**Figure 1 | Interspecific variation in avian plumage colouration.** a–d, Birds display an astonishing variety of colour patterns. The intense blue plumage of male Hainan blue flycatchers (*Cyornis hainanus*) (a) is dramatically more colourful than female plumage (b) in this strongly sexually dichromatic species. In contrast, pale flycatchers (*Bradornis pallidus*) are sexually monochromatic and both males (c) and females (not shown) are drab coloured. Other monochromatic species, however, can be highly ornamented, for example, female red warblers (*Cardellina ruber*) (d) have a degree of plumage colouration that rivals male Hainan blue flycatchers (plumage scores: 71.3 and 71.7, respectively). The evolutionary causes of this diversity are not well understood. Credits: a, S. Kongwittaya; b, M. & P. Wong; c, M. Goodey; d, S. Colenutt.

<sup>1</sup>Institute of Natural & Mathematical Sciences, Massey University, Auckland 0745, New Zealand. <sup>2</sup>Department of Biology, McMaster University, 1280 Main St. West, Hamilton, Ontario L8S 4K1, Canada. <sup>3</sup>School of Biological Sciences, Monash University, Victoria 3800, Australia. <sup>4</sup>Max Planck Institute for Ornithology, Am Obstberg 1, 78315 Radolfzell, Germany. <sup>5</sup>Department of Behavioural Ecology and Evolutionary Genetics, Max Planck Institute for Ornithology, Eberhard Gwinner Str, 82319 Seewiesen, Germany.

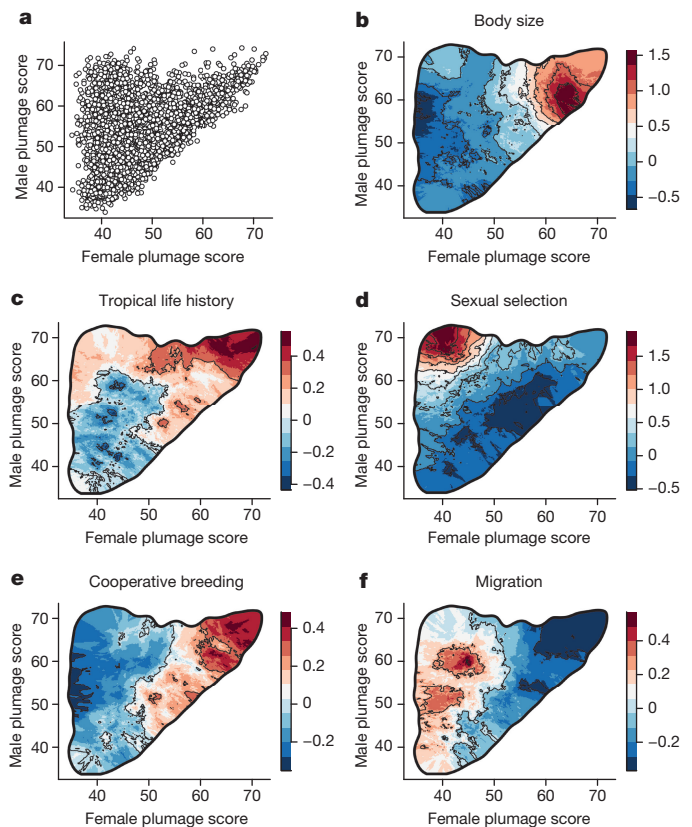


**Figure 2 | The method used for quantifying plumage colouration.**

**a–f.** For each sex of each species we identified similarly coloured plumage patches in all other species and calculated the proportion of those patches that are male. We illustrate this for the male red-billed quelea *Quelea quelea* (image from Lynx Edicions used with permission). **a**, Digitally scanned images of all passerine birds from the *Handbook of the Birds of the World*<sup>18</sup> ( $N = 5,983$ ) were processed in the R package 'colorZapper' (see Methods). **b**, Red, green, blue (RGB) values in three dorsal (nape, crown, forehead) and three ventral (throat, upper breast, lower breast) patches were measured (RGB values for male upper breast shown). **c**, For each plumage patch, colour values for both sexes in all species were pooled. The panel depicts upper breast patch scores visualized in RGB colour space,  $N = 11,966$  points (that is, 2 points for each species, point colour is determined from actual RGB values). **d, e**, For each sex in each species the nearest 1% ( $N = 120$ ) of data points (in Euclidian space) were identified (coloured data points) (d) and the percentage of males (blue points) was determined and used as a 'patch score' (e). **f**, The average of the six patch scores was used as the final 'plumage score' for each sex (male and female red-billed quelea illustrated). Plumage scores thus reflect how 'male-like' or 'female-like' a plumage is, but are determined independently of the sex of each focal data point. There is a high correlation between scores determined as described here versus analogous scores determined with ultraviolet-to-visible reflectance spectra from museum specimens (Extended Data Fig. 1). The overall patterns reported in this paper are highly robust to different cut-offs used to calculate plumage scores (Extended Data Fig. 3).

We quantified male and female plumage colouration in all passerine birds (Order: Passeriformes) illustrated in the *Handbook of the Birds of the World*<sup>18</sup>. Passerines represent the most derived and largest avian radiation ( $N = 5,983$  species, 61% of all birds). Figure 3a illustrates the basic patterns of colour variation in this group: in many species males are more colourful than females, but there are also many sexually monochromatic species with extensive variation in colour elaboration (Extended Data Fig. 2 shows the colours and associated scores for different patches).

Across the passerines, male and female plumage scores were highly correlated (Fig. 3a, phylogenetically controlled reduced major axis



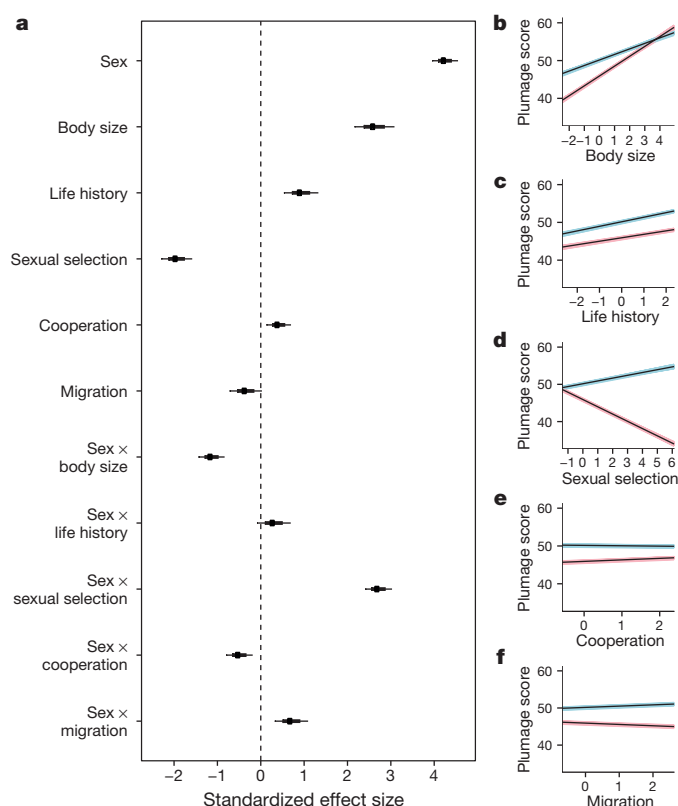
**Figure 3 | Plumage scores and plumage dichromatism in relation to key predictors in passerine birds.** **a**, Male plumage score versus female plumage score ( $N = 5,983$  species). **b–f**, Contour maps depicting the average of the (scaled) predictor values overlaid within the 99.8% volume contour of the male versus female plumage score distribution. Note that the strength of the relationship between predictor values and plumage colouration varies between plots (as reflected by varying ranges in the contour legends). Values within the plots were calculated by superimposing a  $300 \times 300$  grid over the scatter occurring between 30 and 80 and then calculating at each grid point the mean predictor value of the closest 3% of species.

regression,  $R^2 = 0.299$ ,  $T = 5.96$ ,  $P < 0.0001$ ). Multivariate Ornstein–Uhlenbeck evolutionary models indicate that plumage colour evolution is subject to cross-sex constraints that partially restrict independent trait evolution (Extended Data Table 1). Cross-sex constraints, however, do not imply that ornamentation cannot also be directly selected for signalling needs<sup>10</sup>.

To investigate colour diversity among species, we first tested for correlations between sexual dichromatism and ten predictor variables. Sexual dichromatism decreased with body mass and wing length, increased with latitude, more seasonal environments and clutch size, increased with social polygyny, sexual size dimorphism and female-only parental care, decreased with cooperative breeding and increased with migratory behaviour (Extended Data Table 2a).

Because many of the ten predictor variables are strongly intercorrelated, we next consolidated variation to five main predictors (Extended Data Table 3): (1) species with large 'body size' values were heavier and had longer wings; (2) species with high 'tropical life history' values were more likely to breed in the tropics, inhabit areas with year-long environmental stability and lay small clutches; (3) species with high 'sexual selection' scores tended to be socially polygynous, show male-biased sexual size dimorphism and lack paternal care; (4) 'cooperative breeding' was defined as present or absent; and (5) 'migration' as no, partial or complete migration between breeding and non-breeding ranges. Multiple predictor models (Extended Data Table 2b), which control for the confounding effects of the predictors, showed sexual selection to be the strongest predictor of dichromatism (followed by migration



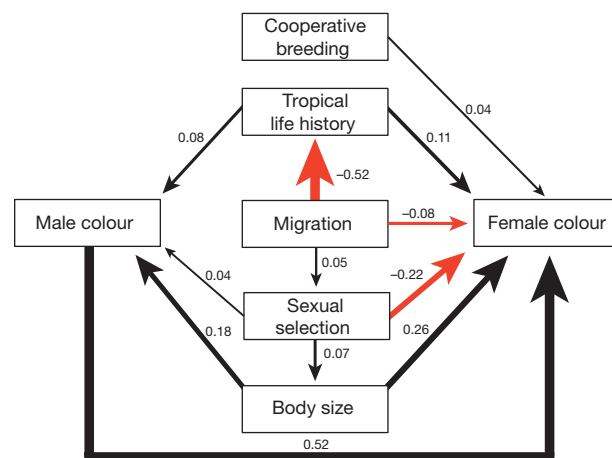


**Figure 4 | Coefficient estimates and model lines of linear mixed models predicting plumage scores in 2,471 species of passerines.** **a**, Coefficients plot of the effect sizes for each of the five key predictor variables (scaled) and their respective interactions with sex. The centre point denotes the mean, the thick bar denotes the posterior standard deviations (akin to the standard errors) and the thin bars denote the 95% lower and upper confidence limits as calculated by model-averaging 100 separate Markov chain Monte Carlo generalized linear mixed models using the package MCMCglmm<sup>20</sup> in R and 100 different phylogenetic trees from <http://birdtree.org><sup>30</sup>. **b–f**, Model predictions of the effect sizes keeping the effects of other predictors constant (males, blue; females, pink). Solid black lines denote the mean effect size of the 100 separate MCMCglmm runs and coloured areas illustrate the full range of regression line estimates across all models.

and body size) while tropical life history did not have a significant effect. These results are difficult to interpret, however, because dichromatism can vary through evolutionary changes in males, females or both sexes<sup>19</sup>. Therefore, to resolve the evolutionary drivers associated with colour elaboration itself, we analysed male and female plumage phenotypes concurrently.

Phylogenetically informed linear mixed models<sup>20</sup> on complete data from 2,471 species (Fig. 4 and Extended Data Table 4) showed that the five predictor variables explained more interspecific variation in female than in male colouration. Female colouration was elaborated (more male-like) in larger species, in species with tropical life histories and in cooperative breeders. In males, larger species and species with tropical life histories also had more colour-elaborated plumage, but the effect of body size was smaller (Fig. 4b). As expected, in species with strong male-biased sexual selection, male colours were significantly more elaborated. However, sexual selection had a much larger and opposite effect on females: in species with strong male-biased sexual selection, females had greatly reduced colour elaboration (Fig. 4d). Plumage dichromatism therefore increased in species with strong sexual selection, as reported in numerous other studies<sup>19</sup>, but our results show that the principal driver of this pattern is evolutionary change in female, not male, colour elaboration.

Because our predictor variables might interact in complex ways, we used phylogenetic path analysis<sup>21</sup> to disentangle cause and effect



**Figure 5 | Relationships among ecological variables and plumage colouration, as determined by phylogenetic controlled *d* separation path analysis<sup>21</sup>.** Arrows indicate direct effects; the strength of the effect is indicated with numeric values and by line thickness. Arrow colour indicates the direction of the effect (black, positive; red, negative).

relationships. This supported the analysis described earlier, demonstrating that sexual selection has a strong direct negative effect on female colouration, and a weaker direct positive effect on male colouration (Fig. 5). Additionally, female colouration was directly influenced by all predictor variables. Models that included a correlated response to male colour were generally favoured, but models in which female colour was only directly influenced by male colour and not by other variables performed poorly (Extended Data Fig. 4 and Extended Data Table 5). Thus, our results suggest that female colouration is not merely a by-product of strong selection on males, but is an adaptive response to various social and life-history factors.

Three main conclusions follow from our study. First, although body size is rarely considered in interspecific studies of animal colouration (but see ref. 22), it strongly predicted colour elaboration in both sexes, and more strongly in females than in males. Larger species are both more colour elaborated and less dichromatic (Fig. 4b), consistent with the hypothesis that being larger reduces predation risk<sup>23</sup>, thereby potentially weakening selection for crypsis. The result refutes the argument that large body size itself is an evolutionary constraint on colouration<sup>22</sup> (at least for passerines).

Second, both sexes of species with tropical life histories, typified by equatorial breeding ranges, low seasonality and small clutch sizes<sup>24</sup>, were more elaborated than temperate breeding species (Fig. 4c), a robust verification of the hypothesis that tropical species are more colourful<sup>25</sup>. The strength of this effect rivals the strength of the effect of sexual selection on colouration (at least for males). These patterns are consistent with two non-mutually exclusive hypotheses. First, colourful plumage functions in mate choice and mutual sexual selection is stronger in tropical species; and second, colourful plumage functions in an aggressive context and resource competition is stronger in tropical species. Indeed, tropical species are thought to be under increased competition for breeding vacancies and resources, have longer-term pair bonds, more common year-round territoriality and increased convergence of male and female reproductive roles<sup>24</sup>. Moreover, there is an obvious convergence between visual and vocal signalling in tropical birds. Female song is more common in tropical species<sup>24</sup> and experimental evidence suggests that song in tropical species has dual functionality as both advertisement for mates as well as an ‘armament’ in competitive interactions<sup>26</sup>.

Third, the intensity of sexual selection strongly predicted variation in plumage colouration (Fig. 4d): more elaborated males and increased sexual dichromatism are found in species with male-biased sexual selection. Increased sexual dichromatism, however, was driven mainly by a strong negative relationship between the intensity of sexual

selection and female colouration. A similar pattern was observed in New World blackbirds (Icteridae)<sup>27</sup>, but our study demonstrates its generality within the passerines. This finding is consistent with two hypotheses. First, it can reflect increased sexual selection on females in monogamous species, supporting a game-theoretic model that predicts stable mutual mate choice in species with extensive parental investment in both sexes<sup>28</sup>. Second, selection for social signalling may be reduced in females of species with strong male-biased sexual selection<sup>27</sup>. Indeed, ecological factors that favour social polygyny<sup>29</sup> (for example, spatially clumped resources) may also be associated with reduced social competition between females, at least in the absence of paternal care. Additionally, high levels of male-biased sexual selection seem to break the correlation between male and female ornamentation (Fig. 4d). Here, the divergent parental roles of males and females may facilitate the evolution of sex-specific developmental modifiers (for example, hormones) that limit the expression of ornamental plumage in females<sup>10</sup>.

Traditionally, studies have focused on male colour elaboration, particularly in species with extreme sexual selection. This has left much of the interspecific variation in colouration unexplained. Our results demonstrate clearly that the immense diversity in avian plumage colouration is the outcome of selection acting additively and often differentially between the sexes (Fig. 3). The patterns presented here for the passerines can be tested in other taxa and provide a rich arena for future hypothesis testing on the function of ornamentation in males and females.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 21 May; accepted 21 August 2015.**

**Published online 4 November 2015.**

- Darwin, C. *The Descent of Man, and Selection in Relation to Sex* (John Murray, 1871).
- Bateman, A. J. Intra-sexual selection in *Drosophila*. *Heredity* **2**, 349–368 (1948).
- Trivers, R. in *Sexual Selection and the Descent of Man 1871–1971* (ed. Campbell, B.) 136–179 (Aldine, 1972).
- Andersson, M. B. *Sexual Selection* (Princeton Univ. Press, 1994).
- Amundsen, T. Why are female birds ornamented? *Trends Ecol. Evol.* **15**, 149–155 (2000).
- Rubenstein, D. R. & Lovette, I. J. Reproductive skew and selection on female ornamentation in social species. *Nature* **462**, 786–789 (2009).
- Tobias, J. A., Montgomerie, R. & Lyon, B. E. The evolution of female ornaments and weaponry: social selection, sexual selection and ecological competition. *Phil. Trans. R. Soc. B* **367**, 2274–2293 (2012).
- Lande, R. in *Sexual Selection: Testing the Alternatives* (eds Bradbury, J. W. & Andersson, M.) 83–94 (Wiley, 1987).
- Bonduriansky, R. & Chenoweth, S. F. Intralocus sexual conflict. *Trends Ecol. Evol.* **24**, 280–288 (2009).
- Kraaijeveld, K. Reversible trait loss: the genetic architecture of female ornaments. *Annu. Rev. Ecol. Syst.* **45**, 159–177 (2014).
- Poissant, J., Wilson, A. J. & Coltman, D. W. Sex-specific genetic variance and the evolution of sexual dimorphism: a systematic review of cross-sex genetic correlations. *Evolution* **64**, 97–107 (2010).
- Potti, J. & Canal, D. Heritability and genetic correlation between the sexes in a songbird sexual ornament. *Heredity* **106**, 945–954 (2011).
- Cardoso, G. C. & Mota, P. G. Evolution of female carotenoid colouration by sexual constraint in Carduelis finches. *BMC Evol. Biol.* **10**, 82 (2010).
- West-Eberhard, M. J. Sexual selection, social competition, and speciation. *Q. Rev. Biol.* **58**, 155–183 (1983).
- Lyon, B. E., Eadie, J. M. & Hamilton, L. D. Parental choice selects for ornamental plumage in American coot chicks. *Nature* **371**, 240–243 (1994).
- West-Eberhard, M. J. Darwin's forgotten idea: the social essence of sexual selection. *Neurosci. Biobehav. Rev.* **46**, 501–508 (2014).
- Lyon, B. E. & Montgomerie, R. Sexual selection is a form of social selection. *Phil. Trans. R. Soc. B* **367**, 2266–2273 (2012).
- del Hoyo, J., Elliott, A. & Christie, D. A. *Handbook of the Birds of the World Vols 8–16* (Lynx Edicions, 2003–2011).
- Badyaev, A. V. & Hill, G. E. Avian sexual dichromatism in relation to phylogeny and ecology. *Annu. Rev. Ecol. Syst.* **34**, 27–49 (2003).
- Hadfield, J. D. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *J. Stat. Softw.* **33**, 1–22 (2010).
- von Hardenberg, A. & Gonzalez-Voyer, A. Disentangling evolutionary cause-effect relationships with phylogenetic confirmatory path analysis. *Evolution* **67**, 378–387 (2013).
- Galván, I., Negro, J. J., Rodríguez, A. & Carrascal, L. M. On showy dwarfs and sober giants: body size as a constraint for the evolution of bird plumage colouration. *Acta Ornithol.* **48**, 65–80 (2013).
- Ricklefs, R. E. Insights from comparative analyses of aging in birds and mammals. *Aging Cell* **9**, 273–284 (2010).
- Stutchbury, B. J. & Morton, E. S. *Behavioral Ecology of Tropical Birds* (Academic, 2001).
- Bailey, S. F. Latitudinal gradients in colors and patterns of passerine birds. *Condor* **80**, 372–381 (1978).
- Tobias, J. A., Gamarra-Toledo, V., García-Olaechea, D., Pulgarín, P. C. & Seddon, N. Year-round resource defence and the evolution of male and female song in suboscine birds: social armaments are mutual ornaments. *J. Evol. Biol.* **24**, 2118–2138 (2011).
- Irwin, R. E. The evolution of plumage dichromatism in the New World blackbirds: social selection on female brightness. *Am. Nat.* **144**, 890–907 (1994).
- Kokko, H. & Johnstone, R. A. Why is mutual mate choice not the norm? Operational sex ratios, sex roles and the evolution of sexually dimorphic and monomorphic signalling. *Phil. Trans. R. Soc. B* **357**, 319–330 (2002).
- Emlen, S. T. & Oring, L. W. Ecology, sexual selection, and the evolution of mating systems. *Science* **197**, 215–223 (1977).
- Jetz, W., Thomas, G. H., Joy, J. B., Hartmann, K. & Mooers, A. O. The global diversity of birds in space and time. *Nature* **491**, 444–448 (2012).

**Acknowledgements** We thank the many ornithologists and scientists who have published their data or contributed data to public databases, allowing us to conduct this study. Thanks to J. D. Aguirre, P. M. Bustin, J. Clavel, P. B. Rainey, L. Redfern and J. A. Tobias for comments on manuscript drafts and to the staff of Museum Victoria and the Australian National Wildlife Collection for access to museum specimens. This work was supported by Massey University and a grant from the Australian and Pacific Science Foundation (APSF 10/8) to J.D. C.J.D. was supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) Canadian Graduate Scholarship. K.D. was supported by the Australian Research Council (DE120102323). B.K. and M.V. were generously supported by the Max Planck Society.

**Author Contributions** Conceived of the study: J.D., M.V. and B.K.; collected the data: J.D., M.V., K.D. and C.J.D.; developed the methods: J.D. and M.V.; analysed the data: J.D. and C.J.D. with help from M.V. and K.D.; wrote the paper: J.D. and C.J.D. with input from the other authors.

**Author Information** Data sets have been deposited in the Dryad Digital Repository (<http://dx.doi.org/10.5061/dryad.1rp0s>). Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.D. ([j.dale@massey.ac.nz](mailto:j.dale@massey.ac.nz)).

## METHODS

**Passerine classification and species list.** Plumage scores were calculated on each of the 5,983 species of passerines (Order: Passeriformes) included in the *Handbook of the Birds of the World* (volumes 8–16)<sup>18</sup>. Phylogenetically controlled statistical analyses were restricted to the 5,831 of these species that are also included in the avian phylogenies at <http://www.birdtree.org><sup>30</sup>. No statistical methods were used to predetermine sample size.

**Plumage scores.** We digitally scanned images of each passerine species from the plates in the *Handbook of the Birds of the World*<sup>18</sup> into 300 dpi JPEGs using a Fuji Xerox ApeosPort-IV C5575 set to default scan settings. Each species was cropped out of the scans and used to measure the RGB (red, green, blue) values for six patches (nape, crown, forehead, throat, upper breast, and lower breast) on each sex of each species using the R package 'colorZapper'<sup>31</sup>. We scored these regions because (1) they are consistently illustrated clearly for each species in the plates (rumps are often not shown for instance), and (2) the anterior body region is unarguably a very important signalling region for birds in general. For each plumage patch, a polygon was subjectively selected that encompassed the typical colouration evident in that area of the bird's plumage (see Fig. 2a for an example with red-billed quelea<sup>32</sup>). We excluded any obvious areas of glare added by the plate artists. colorZapper then calculated the mean values for R, G and B (on scales of 0 to 255) for 400 randomly chosen pixels within the selected polygon. In cases where multiple subspecies were illustrated, we scored colouration in the nominate subspecies.

Males and females are usually illustrated with the same image in the *Handbook of the Birds of the World*<sup>18</sup> when they are sexually monochromatic ( $N = 3,822$  species). In these cases, the same image was measured independently twice, once to obtain male RGB values and once to obtain female RGB values. This was required to maintain similar measuring error between both dichromatic and monochromatic species. We confirmed that within-species measuring error was consistent between species scored with two different images versus species scored with the same image twice. In 63 of the 2,161 species where we used separate images for males and female colour measurements, females differed from males in ways other than the colour of the six plumage patches used in this study (for example, they had shorter tails or a different coloured iris). The correlation between male and female plumage scores from this sample ( $R^2 = 0.905$ ,  $N = 63$ ) was similar to the correlation between male and female plumage scores measured separately from the same image ( $R^2 = 0.913$ ,  $N = 3,822$ ).

Sometimes, when there is only a small difference between male and female plumages, the same image is used to illustrate both sexes of a species and the difference between the sexes is described in the text. In 372 species (6.22% of total) we noted a described difference in the text that was not illustrated in the plates. Many of these descriptions noted that females were "similar to males" but "duller" ( $N = 65$ ), "slightly duller" ( $N = 44$ ), "paler" ( $N = 46$ ), or "slightly paler" ( $N = 29$ ). Most of the rest of the described differences related to specific colour differences (for example, females described as "brownier", "more rufous", "less blue", and so on) or to specific colour patches (for example, "coronal patch absent or small"). To estimate the magnitude of the measurement error associated with this issue, we identified 100 species for which the described difference between the sexes was similar to those described earlier (that is, "duller", "slightly duller", "paler", or "slightly paler") but for which we scored male and female colouration from separate images. In these species, the average difference between male and female plumage scores was 3.82 ( $N = 100$  species,  $P < 0.001$ ).

Because these described differences occur in a small percentage of species and reflect relatively small differences between the sexes, we expected them only negligibly to affect our general results. To confirm that our main results are robust to the error generated by these species, we repeated the main MCMCglmm analysis in two different ways: (1) with these species removed entirely from the data set (note that these species affected 168 of the 2,471 species (6.8%) in the MCMCglmm); and (2) with these species' plumage scores adjusted by subtracting the mean male–female difference calculated as described earlier (that is, 3.82) from the female scores. Both of these analyses yielded essentially identical results to the analysis reported in the paper (the  $R^2$  between the effects of these analyses and the effects reported in Extended Data Table 4 was 0.998 and 0.999, respectively).

Scoring colouration with handbook plates represents a valid alternative to measuring colour on all the world's passerines using museum specimens and/or live individuals. Indeed, the objective of handbook plates is to reflect as accurately as possible the typical colouration and patterning of a species such that the image is a suitable reference for field identification. As such, special care is taken to reproduce colouration accurately and the plates in the *Handbooks of the Birds of the World* are highly regarded as superb in quality and consistency (for example, see ref. 33). Moreover, previous studies have found that colouration in plates is highly correlated with colouration in museum specimens (as measured by spectrometry<sup>34</sup>), and several previous comparative studies have used colouration in plates to test hypotheses about avian colour evolution (for example, see refs 35–38). Finally,

we validated that plumage scores generated with handbook plates were highly correlated with plumage scores generated with ultraviolet-to-visible (UV–Vis) spectrometry (see later and Extended Data Fig. 1).

**Plumage scores validation analysis.** We validated that plumage scores are consistent between different colour-measuring methodologies. We used UV–Vis spectrometry to measure the reflectance of 8 plumage patches (upper back, dorsal neck, crown, forehead, throat, ventral neck, upper breast and lower breast) of up to 3 male and 3 female museum specimens for 534 species of Australian terrestrial birds (229 non-passerines and 305 passerines). Museum specimens were obtained from Melbourne Museum (Museum Victoria) and the Australian National Wildlife Collection (CSIRO). Reflectance spectra were collected using a spectrometer (Avaspec 2048, Avantes) connected to a xenon pulsed light source (Avalight-XE) through a fibre optics cable fitted at the end with a plastic cylinder to exclude ambient light and standardize measuring distance. Reflectance spectra were expressed relative to a WS-2 white standard using Avasoft software. For each species we calculated mean reflectance spectra per patch separately for males and females. To compute plumage scores we pooled together all male and female mean spectra for each patch ( $N = 1,068$  spectra for each patch). We matched each spectrum in the pool to the 120 most similar spectra as determined by minimizing the squared-differences between spectra at 5 nm wavelength intervals, summed across 300–700 nm. We then quantified the percentage of those 120 'closest matches' that corresponded to male spectra. For each species, male and female UV–Vis plumage scores were calculated as the mean 'percentage male' values across all of the eight patches measured in each sex of each species. For these same 534 species we also quantified patch colour using handbook plates digitally scanned from those published previously<sup>18</sup> and then scoring them in RGB colour space using the R-package colorZapper<sup>31</sup> (see earlier and Fig. 2). The 120 closest matching colour patches to any focal patch were determined by minimizing the Euclidian distance in RGB space, and then the proportion males in those closest matches was quantified. Note that this method of quantifying colouration is sample-size dependent: larger samples of species will provide more accurate measures of how 'male-like' any kind of colour is. Nevertheless, despite having only 534 species in this sample (contrasting with 5,983 species in the main analysis), both 'male-like' indices were strongly positively correlated ( $R^2 = 0.67$ ,  $P < 0.0001$ ; Extended Data Fig. 1a), indicating that using handbook plates to estimate bird colours is a suitable alternative to the use of reflectance spectrometry.

To demonstrate further that plumage scores calculated with handbook plate measurements are suitably interchangeable with plumage scores calculated with UV–Vis spectra, we repeated statistical analyses presented in the paper with the subset of passerine species for which spectral data were collected. If handbook plates provide a suitable surrogate measure then conclusions drawn from spectral data on the 305 species should be similar to conclusions drawn from the same 305 species based on plate data. First, we compared effect size estimates for single predictor PGLS models run on each of the ten predictor traits in Extended Data Table 2a. We ran models predicting female scores, male scores and dichromatism scores ( $N = 30$  effects in total). Our analysis showed a very strong correlation between RGB effect sizes versus UV–Vis effect sizes (Extended Data Fig. 1b). Second, we compared effect size estimates (including sex–predictor interactions) for MCMCglmm models run with our five main predictors in Fig. 4. For each plumage score type we ran five models using a different phylogenetic tree<sup>30</sup> in each model. This analysis showed a very strong correlation between RGB effect sizes versus UV–Vis effect sizes (Extended Data Fig. 1c). In combination, these analyses strongly indicate that the biological patterns reported in the main analysis would not have been different if we had used plumage scores based on UV–Vis spectra. These results thus provide critical validation of our method because although human and avian vision have considerable overlap<sup>19,39</sup>, birds can also see UV light not visible to humans<sup>40</sup>.

**Predictor variables.** Body size (mass and wing length) was tested because it is a known confound of sexual size dimorphism<sup>41</sup> and life-history traits<sup>42</sup>. In addition, based on a previous study<sup>22</sup> we predicted a negative relationship between body size and colour elaboration. Tropical life history (latitude, seasonality and clutch size) was tested to evaluate the prediction that characteristics associated with tropical breeding—in particular, density-dependent or 'K-selected'<sup>43</sup> factors such as increased competition for limited breeding vacancies<sup>24</sup>—were associated with increased plumage colour elaboration. Sexual selection (social mating system, sexual size dimorphism, and paternal care) was tested because the prevailing view in the literature is that sexual selection on males is the fundamental evolutionary driver of male colour elaboration and sexual dimorphism. However, recent studies on New World blackbirds<sup>27,44,45</sup> (Icteridae), tanagers<sup>46</sup> (Thraupidae) and fairy-wrens<sup>47</sup> (Maluridae) have demonstrated rapid evolutionary transitions in female colouration (also see ref. 48). Therefore we also expected sexual selection on males to potentially have a strong effect on female colouration. Cooperative breeding species were predicted to be more colour elaborated because of increased social



competition among group members and this effect was predicted to be stronger in females than in males<sup>6,49</sup>. Migration was tested to evaluate two opposing predictions. On the one hand, it has been argued<sup>50</sup> that species with long-distance migration should be less colourful as a result of increased predation associated with long-distance movements (also see ref. 51). In contrast, other researchers<sup>19,52</sup> have argued that migratory species should be more colourful because migration imposes a short mate sampling period and enhances selection on signals related to migration ability.

Body mass data was taken from that described previously<sup>53</sup>. When more than one body mass entry was available for a species, we computed the mean weighted by sample size. Log-transformed values were used in the statistical analysis.

Wing length data was taken from that described previously<sup>41</sup>. Up to 7 different sets (average = 2.5) of wing measurements were recorded per species. Each set comprised the means (or mid-ranges when only a range was provided) for both males and females measured in a single population, and we took the means of these for final species values. Log-transformed values were used in the statistical analysis.

For latitude data, species' geographical location was computed based on the breeding range maps of all species<sup>54</sup>. First, breeding range polygons were transformed from an unprojected coordinate system (latitude, longitude) to an equal-area projected coordinate system (cylindrical equal-area, latitude of the origin: 0°). Species' geographical location was then computed as the latitude (degrees from equator) of the breeding range centroid.

For seasonality, Moderate Resolution Imaging Spectroradiometer (MODIS) land surface temperature rasters (code name MOD11C2) were obtained through the <http://reverb.echo.nasa.gov> gateway at a 0.05° spatial resolution and an 8-day temporal resolution (time span: 2000–2012)<sup>55</sup>. All the raster files were then superposed and a temperature time series was obtained for each pixel. A coefficient of variation (CV%) of temperature was computed for each pixel<sup>54</sup>. For each species the CV% for all pixels within its breeding range were extracted. A species 'seasonality' score was defined as the median CV% score within its breeding range (log transformed).

Clutch size was compiled from standard references, in particular<sup>18,56–60</sup>. Up to five different reports of clutch sizes were recorded for each species, with the mean of these taken as the final clutch size of a species. Clutch size was recorded as the mean clutch size or, when only range data was provided, as the mid-range value. Log-transformed values were used in the statistical analysis.

Sexual size dimorphism was calculated as  $\log(\text{male wing length}) - \log(\text{female wing length})$ <sup>41</sup>, providing a proportional index of relative sizes of the sexes. Positive values reflect species where males are larger than females.

Social polygyny was scored on a four-point scale<sup>36</sup>, with 0 = strict social monogamy (for example, zebra finch *Taeniopygia guttata*), 1 = monogamy with infrequent instances of polygyny observed (<5% of males, for example, lazuli bunting *Passerina amoena*), 2 = mostly social monogamy with regular occurrences of facultative social polygyny (5 to 20% of males, for example, American redstart *Setophaga ruticilla*), and 3 = obligate resource defence polygyny (>20% of males, for example, red-winged blackbird *Agelaius phoeniceus*) or lek polygyny (for example, lance-tailed manakin *Chiroxiphia lanceolata*). Assignments were made based on standard references, in particular<sup>18,41,56–62</sup>. There are a small number of passerine species with polygynandrous mating systems (for example, the dunnoek *Prunella modularis*, Smith's longspur *Calcarius pictus* and sickle-billed vanga *Falcula palliata*) and these species were pooled with the monogamous species. We reasoned that sexual selection would be more similar in each sex in polygynandrous species compared with polygynous species, and our social polygyny scores were intended to specifically quantify male-biased sexual selection.

Paternal care was scored as absent (0) or present (1) primarily based on the data provided in ref. 63, including both the known and inferred data categories. For species in our data set not present in ref. 63, we used standard references, in particular ref. 18, to obtain the additional parental care scores.

Cooperative breeding was scored as absent (0), suspected (0.5), or present (1) primarily based on the data provided in ref. 63, including both the known and inferred data categories. For species in our data set not present in ref. 63, we used standard references, in particular ref. 18, to obtain the additional cooperative breeding scores.

Migration was scored on a scale from 0 to 2, with 0 = resident (that is, breeding and non-breeding ranges identical), 1 = partial migration (that is, some overlap between breeding and non-breeding ranges), 2 = complete migration (that is, no overlap between breeding and non-breeding ranges). Assignments were made based on the range maps published previously<sup>18</sup>. The migratory behaviour of one species, the Red Sea Swallow *Hirundo perditia*, is unknown.

**Statistical analysis.** Because closely related species tend to be more similar to each other than distantly related species, we used phylogenetically informed comparative analyses to account for potential non-independence among species owing to common ancestry. More specifically, the error structure of the statistical model

incorporates the degree of non-independence between species as estimated from the phylogeny. Phylogenetically informed methods are unlike ordinary statistical models (where data points are assumed to be independent) because they explicitly model how the covariance between species declines as they become more distantly related<sup>64–66</sup>. We used the Hackett<sup>67</sup> backbone phylogenetic trees available at <http://birdtree.org><sup>30</sup> to estimate phylogenetic separation in our statistical models.

The relationships between the ten predictor variables listed earlier and sexual dichromatism (Extended Data Table 2) were modelled with phylogenetic generalized least-squares<sup>64–66,68</sup> (PGLS) using the R-packages 'ape'<sup>69</sup> and 'nlme'<sup>70</sup>.

To calculate phylogenetic reduced major axis regression<sup>71</sup> and the phylogenetic principal components<sup>72</sup> used in the multiple predictor PGLS, MCMCglmm and paths analysis we used the R-package 'phytools'<sup>73</sup>.

All data were analysed in R<sup>74</sup> version 3.1.0. To improve the interpretability of regression coefficients<sup>75</sup>, predictor variables were centred and standardized to a mean = 0 and standard deviation = 1.

**Phylogenetically informed generalized linear mixed models.** Monte Carlo Markov chain generalized linear mixed models (Fig. 4 and Extended Data Table 4) were generated with the R-package 'MCMCglmm'<sup>20</sup>. MCMCglmm is a Markov chain Monte Carlo sampler for multivariate mixed models that enables the inclusion of a phylogeny as a design matrix in a Bayesian generalized linear modelling framework. The design matrix for phylogenetic effects was based on 100 trees from <http://birdtree.org><sup>30</sup>. We then used MCMCglmm to fit male and female plumage scores as the response and the five predictor variables as continuous predictors. Phylogenetic effects were considered a random effect, sex was fit as a dummy variable and species was fit as an observation level random effect. We used the prior:  $[list(R = list(V = 1, nu = 0.002), G = list(G1 = list(V = 1, nu = 0.002)))]$  and model outcomes were insensitive to prior parameterization. We let the MCMC algorithm run for 10,000,000 iterations, with a burn in period of 3,000 and a sampling interval of 10,000. Each model generated ~1,000 independent samples of model parameters (Extended Data Table 4). Independence of samples in the Markov chain was assessed by graphic diagnostics and testing for autocorrelation between samples.

**Multivariate Ornstein–Uhlenbeck evolutionary models.** Ornstein–Uhlenbeck (OU) models are powerful tools for analysing the evolution of traits, and are well suited to studies of correlated evolution<sup>76</sup>. These models are a generalization of Brownian motion (BM) models<sup>77</sup>, in that they incorporate stochastic trait evolution (defined by a drift parameter  $\sigma$ ), but also allow trait values to be attracted towards optima, at a rate that is dependent on the value of  $\alpha$  (the strength of selection towards the optima). When  $\alpha = 0$ , the process reduces to Brownian motion; however, many patterns of trait evolution are better fit by OU models, because trait evolution is often subject to stabilizing selection, at least over evolutionary time scales<sup>78</sup>. Additionally, the development of multivariate OU (mvOU) models has allowed trait evolution to be influenced by interactions with other trait values. Indeed, trait evolution is often thought of as a multivariate process, and mvOU models allow researchers to explore the influence of co-adaptive or limiting forces on trait evolution.

We used mvOU models to test whether the evolution of plumage ornamentation was subject to cross-sex constraints. Using the mvMORPH<sup>79</sup> package in R, we constructed five evolutionary models of the potential relationship between male and female ornamentation (using the plumage scores described in the main text ( $N = 5,831$  species)) and compared their fit with AIC<sup>76</sup>. In model 1, male and female ornamentation evolved completely independently (that is, both the  $\alpha$  and  $\sigma$  matrix were restricted to be diagonal) under OU processes, simulating no evolutionary relationship between male and female plumage colour. In model 2, male and female plumage scores evolved independently towards their optima (a diagonal  $\alpha$  matrix); however, we allowed covariation in the stochastic element (that is, drift) of the OU process. This model might reflect a situation where male and female ornamentation respond similarly to some environmental parameter, but are not directly constrained by one another. In model 3, the evolution of male and female plumage ornamentation was influenced by an interaction with the other sex's trait value (a symmetric positive  $\alpha$  matrix), and we allowed covariation in the stochastic element of trait evolution. Such a model represents cross-sex constraints in plumage colour evolution. Models 4 and 5 were BM models including independent and covarying drift matrices, respectively.

Model 3 was the clear best model (Extended Data Table 1a), demonstrating that male and female plumage colour do not evolve independently. Positive off-diagonal elements in the  $\alpha$  matrix of this model (Extended Data Table 1b) suggest that male and female plumage values are pulled towards one another, consistent with cross-sex constraints against independent trait evolution, and the correlated response hypothesis. Models of fully independent trait evolution performed poorly (model 1 and 4), and models including covariation in drift performed better than those with independent drift. This suggests that in addition to an interaction in selection on male and female plumage colour, the sexes generally have a correlated response to ecological conditions.

**Phylogenetic path analysis.** Confirmatory path analysis is a special case of structural equation modelling used to build models of causal relationships among a suite of variables and test how the data conform with those causal models<sup>80,81</sup>. When conducted in a phylogenetic framework<sup>21,82</sup>, this approach can infer the most likely evolutionary pathways by simultaneously considering both indirect and direct effects among variables, as well as the magnitude of these effects. This technique is advantageous in that it better reflects the true nature of evolutionary processes because factors do interact in complex ways (for example, by acting as both causal parents and causal children) to create evolutionary outcomes (that is, phenotypes). Here, we used the *d*-separation<sup>21,80,82</sup> method to test for the most likely relationships among seven variables related to plumage colouration. The variables used were the five principal components derived from the phylogenetic principle components analysis (Extended Data Table 3), as well as male plumage score and female plumage score. First, we built 14 biologically relevant models of the relationships among these variables (Extended Data Fig. 4). These models are constructed as directed acyclic graphs, which are required for this type of analysis. Model construction was based on the hypothesized relationships among variables that had been suggested in previous studies (for example, see refs 19, 27, 28) and from exploratory analysis of correlations among variables. Each model is then converted to a set of conditional independencies of the form (X1, X2) {X3}, where variables X1 and X2 are independent conditional on variable X3 (see elsewhere<sup>80,82</sup> for details). We then tested the set of conditional independencies in each model using phylogenetic generalized least squares (GLS) models (using the nlme<sup>70</sup> package in R<sup>74</sup>) and a single tree that was randomly selected from the tree pool at <http://birdtree.org><sup>30</sup>. Using these GLS models, we calculated Fisher's *C*-statistic (which is equivalent to a maximum-likelihood estimate<sup>83</sup>) and the *C*-statistic information criterion (CICc) for each conceptual model<sup>82,83</sup>.

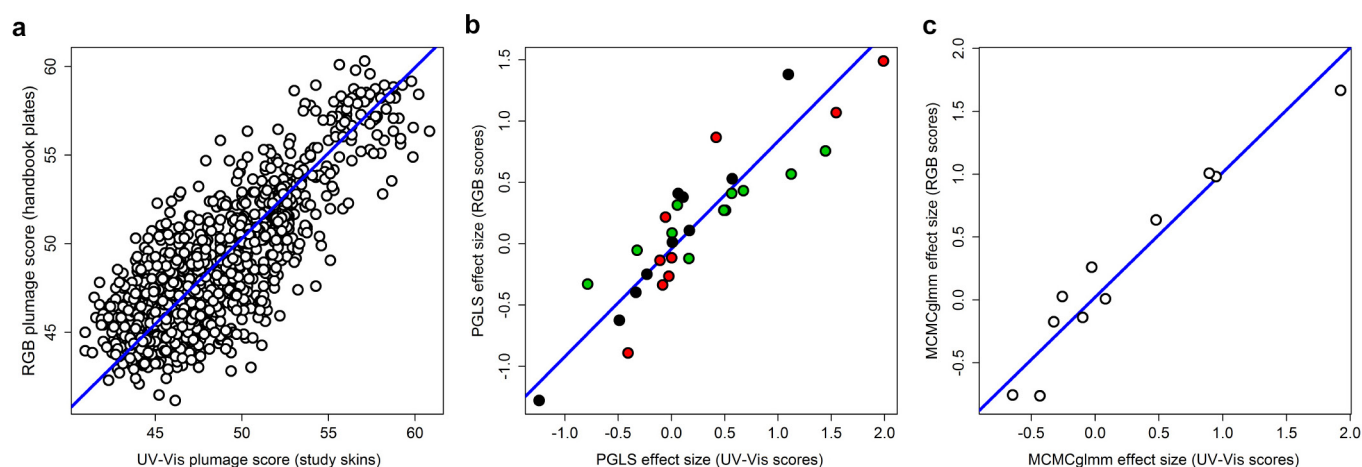
Comparison of models based on CICc values indicated that model K was the best model (Extended Data Table 5). Model L appears to also be a competitive model, however, visual comparison of models K and L show that they suggest the same causal relationships among variables, except model L contains one additional causal link (between life history and sexual selection). The addition of a single, uninformative variable often causes marginal changes in CIC and can therefore appear to create competitive models, although in such a case the simpler model should be preferred<sup>84,85</sup>. Models including a direct effect of male colour on female colour performed better than those that did not include this effect (G, I) and those in which there was a direct effect of female colour on male colour (M, N). However, a model in which female colour was only directly influenced by male colour and not by other variables (C) also performed poorly. This suggests that female colouration is not simply a genetically correlated response to selection on male colouration, and is instead affected by other variables. The best model (K) generally supports the results of the MCMCglmm analysis (Fig. 4 and Extended Data Table 4). This model (Fig. 5) suggests that body size, cooperative breeding, migration, tropical life history, sexual selection and male colour all directly influence female colouration, while only sexual selection, tropical life history and body size directly influence male colouration. The magnitude and direction of the standardized regression coefficients (Fig. 5) suggest that there is a strong role of social selection on female colouration. Females become more colourful with increasing body size, more tropical life histories and cooperative breeding, factors that are predicted to increase the intensity of competition (see main text). Also consistent with our MCMCglmm analysis, the path analysis shows that sexual selection has direct, but antagonistic effects on both male and female colouration. Interestingly, the negative effect of sexual selection on female colouration is larger than the positive effect on male colouration. This pattern supports the finding that sexual selection has an overall negative effect on passerine colouration.

**Code availability.** Scripts of analyses and code used for figure production are available upon request from J.D.

31. Valcu, M. & Dale, J. colorZapper: color extraction utilities. R package version 1.0. <https://github.com/valcu/colorZapper> (2014).
32. Craig, A. in *Handbook of the Birds of the World* Vol. 15 (eds Del Hoyo, J., Elliot, A. & Christie, D. A.) (Lynx Edicions, 2010).
33. Starck, J. M. Review of *Handbook of the Birds of the World*. *Ethology* **102**, 436–440 (1996).
34. Badyaev, A. V. & Hill, G. E. Evolution of sexual dichromatism: contribution of carotenoid-versus melanin-based colouration. *Biol. J. Linn. Soc.* **69**, 153–172 (2000).
35. Gray, D. A. Carotenoids and sexual dichromatism in North American passerine birds. *Am. Nat.* **148**, 453–480 (1996).
36. Owens, I. P. F. & Hartley, I. R. Sexual dimorphism in birds: why are there so many different forms of dimorphism? *Proc. R. Soc. Lond. B* **265**, 397–407 (1998).
37. Olson, V. A. & Owens, I. P. F. Interspecific variation in the use of carotenoid-based colouration in birds: diet, life history and phylogeny. *J. Evol. Biol.* **18**, 1534–1546 (2005).
38. Dey, C. J., Valcu, M., Kempenaers, B. & Dale, J. Carotenoid-based bill coloration functions as a social, not sexual, signal in songbirds (Aves: Passeriformes). *J. Evol. Biol.* **28**, 250–258 (2015).
39. Seddon, N., Tobias, J. A., Eaton, M. & Odeen, A. Human vision can provide a valid proxy for avian perception of sexual dichromatism. *Auk* **127**, 283–292 (2010).
40. Cuthill, I. C. in *Bird Colouration, Volume 1: Mechanisms and Measurements* (eds Hill, G. E. & McGraw, K. J.) 3–40 (Harvard Univ. Press, 2006).
41. Dale, J. *et al.* Sexual selection explains Rensch's rule of allometry for sexual size dimorphism. *Proc. R. Soc. B* **274**, 2971–2979 (2007).
42. Calder, W. A. *Size, Function, and Life History* (Courier Corporation, 1996).
43. Pianka, E. R. On *r*- and *K*-selection. *Am. Nat.* **104**, 592–597 (1970).
44. Hofmann, C. M., Cronin, T. W. & Orland, K. E. Evolution of sexual dichromatism. 1. Convergent losses of elaborate female colouration in New World orioles (*Icterus* spp.). *Auk* **125**, 778–789 (2008).
45. Price, J. J. & Eaton, M. D. Reconstructing the evolution of sexual dichromatism: current color diversity does not reflect past rates of male and female change. *Evolution* **68**, 2026–2037 (2014).
46. Burns, K. J. A phylogenetic perspective on the evolution of sexual dichromatism in tanagers (Thraupidae): the role of female versus male plumage. *Evolution* **52**, 1219–1224 (1998).
47. Johnson, A. E., Jordan Price, J. & Pruett-Jones, S. Different modes of evolution in males and females generate dichromatism in fairy-wrens (Maluridae). *Ecol. Evol.* **3**, 3030–3046 (2013).
48. Dunn, P. O., Armenta, J. K. & Whittingham, L. A. Natural and sexual selection act on different axes of variation in avian plumage color. *Sci. Adv.* **1**, e1400155 (2015).
49. Rubenstein, D. R. Sexual and social competition: broadening perspectives by defining female roles. *Phil. Trans. R. Soc. B* **367**, 2248–2252 (2012).
50. Alerstam, T., Hedenström, A. & Åkesson, S. Long-distance migration: evolution and determinants. *Oikos* **103**, 247–260 (2003).
51. Simpson, R. K., Johnson, M. A. & Murphy, T. G. Migration and the evolution of sexual dichromatism: evolutionary loss of female colouration with migration among wood-warblers. *Proc. R. Soc. B* **282**, 20150375 (2015).
52. Fitzpatrick, S. Colourful migratory birds: evidence for a mechanism other than parasite resistance for the maintenance of 'good genes' sexual selection. *Proc. R. Soc. Lond. B* **257**, 155–160 (1994).
53. Dunning, J. B. *CRC Handbook of Avian Body Masses* 2nd edn (CRC, 2008).
54. Valcu, M., Dale, J. & Kempenaers, B. rangeMapper: a platform for the study of macroecology of life-history traits. *Glob. Ecol. Biogeogr.* **21**, 945–951 (2012).
55. Land Processes Distributed Active Archive Center (LP DAAC). *MODIS/Terra Land Surface Temperature/Emissivity 8-Day L3 Global 0.05Deg* (LP DAAC, 2014).
56. Marchant, S. & Higgins, P. J. *Handbook of Australian, New Zealand & Antarctic Birds* (Oxford Univ. Press, 1990–2006).
57. Cramp, S. & Simmons, K. E. L. *Handbook of the Birds of Europe, the Middle East and North Africa: the Birds of the Western Palearctic* (Oxford Univ. Press, 1977–1994).
58. Brown, L. H., Urban, E. K. & Newmann, K. *The Birds of Africa* (Academic, 1982–2004).
59. Hockey, P. A. R., Dean, W. R. J. & Ryan, P. G. *Birds of Southern Africa* 7th edn (John Voelcker Bird Book Fund, 2005).
60. Poole, A. & Gill, F. *Birds of North America* (Cornell Lab of Ornithology, 1992–2003).
61. Dunn, P. O., Whittingham, L. A. & Pitcher, T. E. Mating systems, sperm competition, and the evolution of sexual dimorphism in birds. *Evolution* **55**, 161–175 (2001).
62. Pitcher, T. E., Dunn, P. O. & Whittingham, L. A. Sperm competition and the evolution of testes size in birds. *J. Evol. Biol.* **18**, 557–567 (2005).
63. Cockburn, A. Prevalence of different forms of parental care in birds. *Proc. R. Soc. B* **273**, 1375–1383 (2006).
64. Martins, E. P. & Hansen, T. F. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *Am. Nat.* **149**, 646–667 (1997).
65. Freckleton, R. P., Harvey, P. H. & Pagel, M. Phylogenetic analysis and comparative data: a test and review of evidence. *Am. Nat.* **160**, 712–726 (2002).
66. Pagel, M. Inferring the historical patterns of biological evolution. *Nature* **401**, 877–884 (1999).
67. Hackett, S. J. *et al.* A phylogenomic study of birds reveals their evolutionary history. *Science* **320**, 1763–1768 (2008).
68. Garland, T. Jr & Ives, A. R. Using the past to predict the present: confidence intervals for regression equations in phylogenetic comparative methods. *Am. Nat.* **155**, 346–364 (2000).
69. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
70. Pinheiro, J., Bates, D., DebRoy, S. & Sarkar, D. nlme: Linear and nonlinear mixed effects models. R package version 3.1–117. <http://CRAN.R-project.org/package=nlme> (2014).
71. Revell, L. J. Phylogenetic signal and linear regression on species data. *Methods Ecol. Evol.* **1**, 319–329 (2010).
72. Revell, L. J. Size-correction and principal components for interspecific comparative studies. *Evolution* **63**, 3258–3268 (2009).
73. Revell, L. J. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3**, 217–223 (2012).

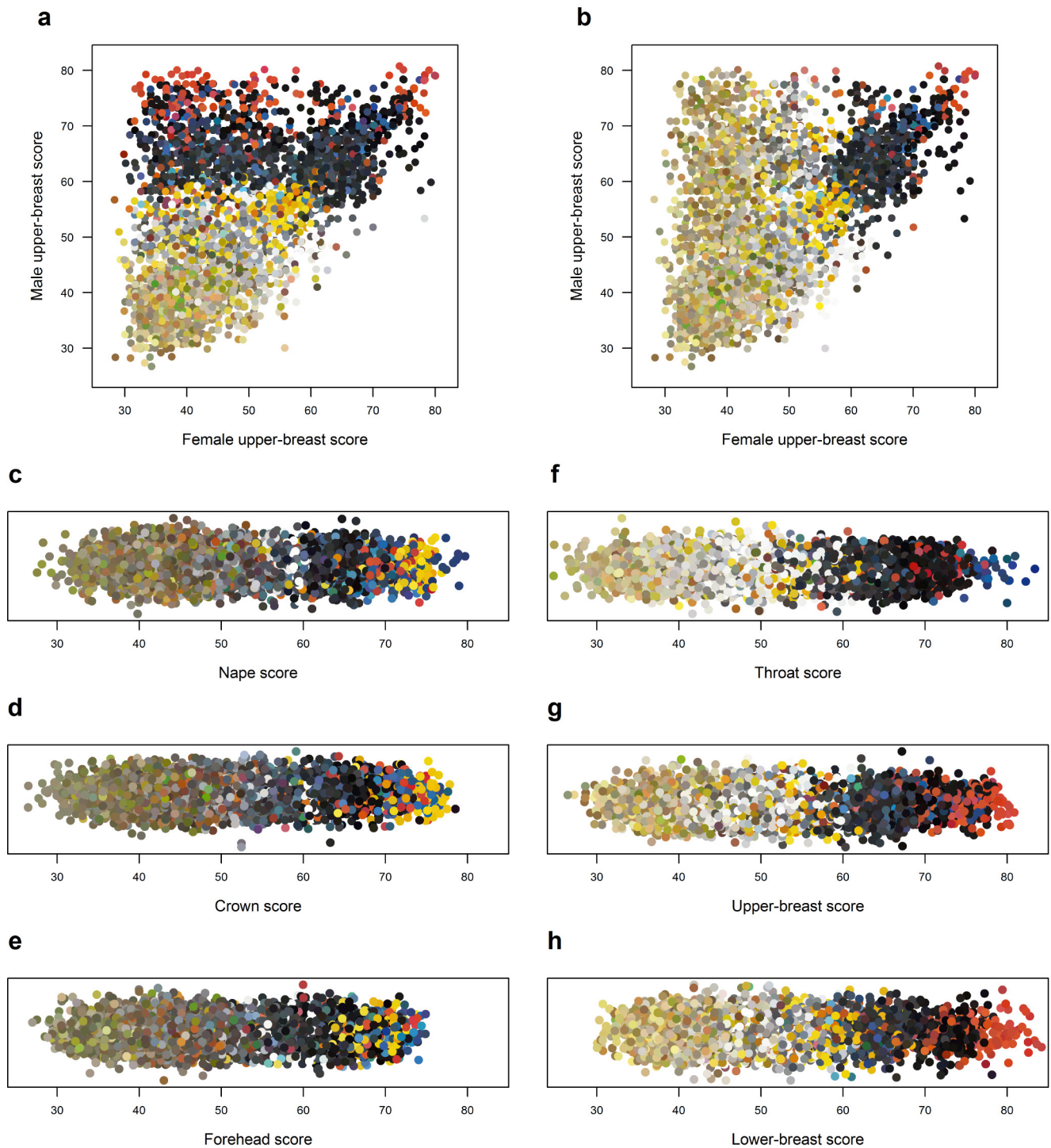
74. R Core Team. *R: A language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2014).
75. Schielzeth, H. Simple means to improve the interpretability of regression coefficients. *Methods Ecol. Evol.* **1**, 103–113 (2010).
76. Bartoszek, K., Pienaar, J., Mostad, P., Andersson, S. & Hansen, T. F. A phylogenetic comparative method for studying multivariate adaptation. *J. Theor. Biol.* **314**, 204–215 (2012).
77. Beaulieu, J. M., Jhwueng, D. C., Boettiger, C. & O'Meara, B. C. Modeling stabilizing selection: expanding the Ornstein-Uhlenbeck model of adaptive evolution. *Evolution* **66**, 2369–2383 (2012).
78. Hansen, T. F. in *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology* (ed. Garamszegi, L. Z.) 351–379 (Springer, 2014).
79. Clavel, J., Escarguel, G. & Merceron, G. mvMORPH: an R package for fitting multivariate evolutionary models to morphometric data. *Meth. Ecol. Evol.* <http://dx.doi.org/10.1111/2041-210X.12420> (2015).
80. Shipley, B. A new inferential test for path models based on directed acyclic graphs. *Struct. Equ. Modeling* **7**, 206–218 (2000).
81. Shipley, B. *Cause and Correlation in Biology: a User's Guide to Path Analysis, Structural Equations and Causal Inference* (Cambridge Univ. Press, 2002).
82. Gonzalez-Voyer, A. & von Hardenberg, A. in *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology* (ed. Garamszegi, L. Z.) 201–229 (Springer, 2014).
83. Shipley, B. The AIC model selection method applied to path analytic models compared using a d-separation test. *Ecology* **94**, 560–564 (2013).
84. Arnold, T. W. Uninformative parameters and model selection using Akaike's information criterion. *J. Wildl. Mgmt.* **74**, 1175–1178 (2010).
85. Burnham, K. P. & Anderson, D. R. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (Springer, 2002).





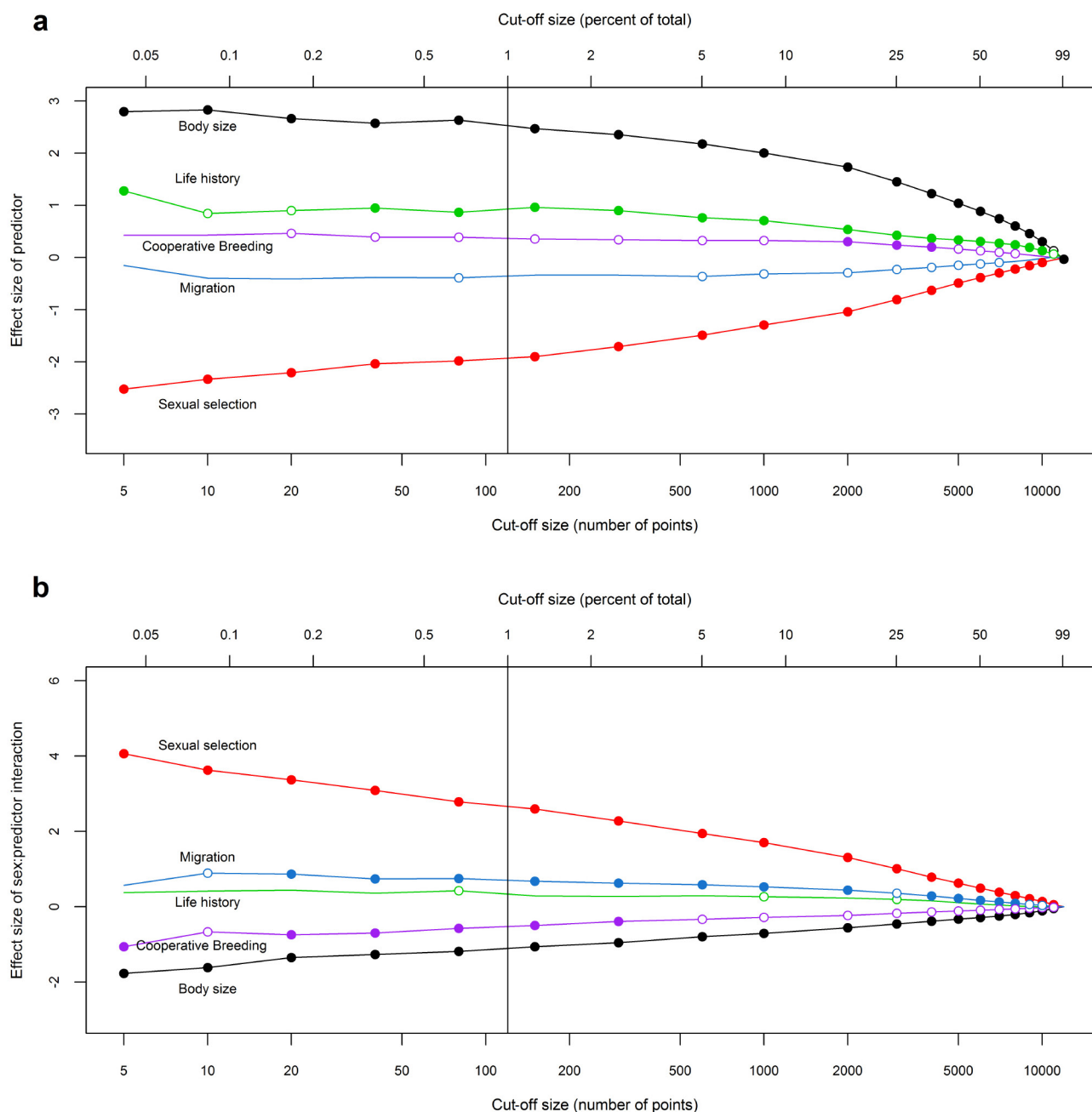
**Extended Data Figure 1 | Comparison of plumage scores determined with handbook plates in RGB colour space with plumage scores determined with study skins using UV-Vis spectrometry.** **a**, RGB versus UV-Vis scores calculated with 534 Australian bird species (reduced major axis (RMA) regression:  $y = 0.965x + 2.025$ ,  $N = 1068$ ,  $R^2 = 0.670$ ,  $P < 0.0001$ ). **b**, PGLS model effect sizes determined with RGB scores versus effect sizes determined with UV-Vis scores (RMA regression:  $y = 0.878x - 0.042$ ,  $N = 30$  model effects,  $R^2 = 0.809$ ,  $P < 0.0001$ , each

model had 305 Australian passerine species in it and black, red and green points reflect models predicting female scores, male scores and dichromatism scores, respectively). **c**, MCMCglmm model effect sizes (including interaction effects) determined with RGB versus UV-Vis scores (RMA regression:  $y = 0.991x - 0.028$ ,  $R^2 = 0.930$ ,  $P < 0.0001$ ,  $N = 11$  effects, the model had 305 Australian passerine species in it and each point reflects the mean effect size calculated from five models using a separate phylogenetic tree from <http://birdtree.org> each).



**Extended Data Figure 2 | Basic patterns of plumage colouration in the order Passeriformes ( $N = 5,983$  species).** **a**, Male versus female patch scores (upper breast shown as an example). Data points are coloured by the male RGB values scored from handbook plates. **b**, As in **a**, only points are coloured with female RGB scores. **c–h**, The colours associated with different plumage scores differentiated by patch type. The y axis is

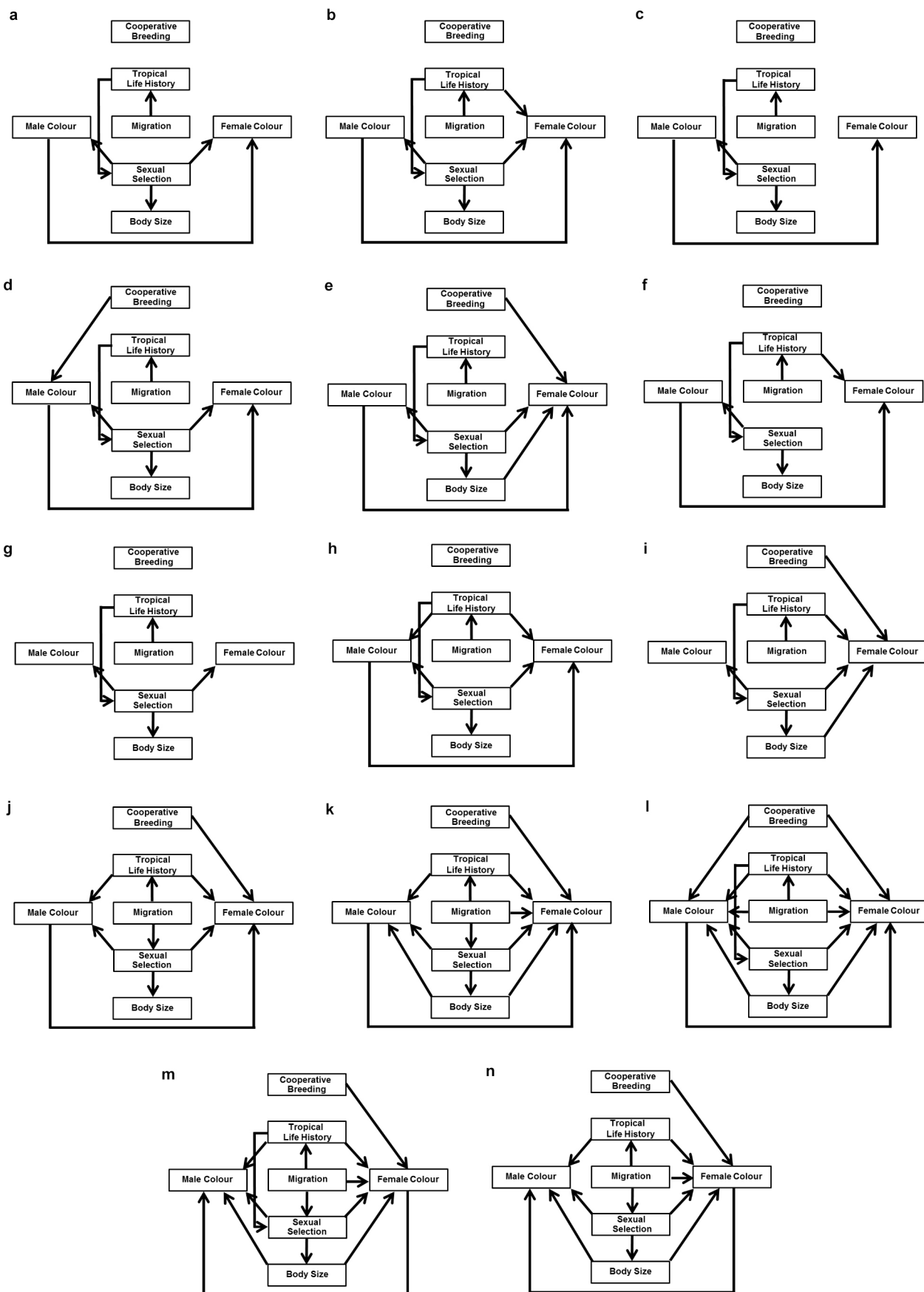
a normally distributed random number used to spread out variation and improve visualization. The figure reflects patterns that agree with our intuition: dull greens, olives and browns have low plumage scores (are female-like), while richer or high contrasting colours (blacks, purples, blues, reds and yellows) have high scores (are male-like).



**Extended Data Figure 3 | MCMCglmm results are robust to various cut-offs used to determine plumage scores. a, b, Main effects (a) and interaction (b) with sex effects. Filled circles: estimates where  $P < 0.001$ ; open circles: estimates where  $0.05 > P > 0.001$ . The vertical line represents the cut-off value used in the main analysis. To reduce processing time for this analysis each MCMCglmm model ran for 10,000 iterations with a sampling interval of 100. Parameter estimates from the shortened runs**

were highly similar to the parameter estimates reported in the main MCMCglmm analysis (see Extended Data Table 4). Note that the observed decline in effect sizes as the cut-off size increases is the automatic outcome of how plumage scores are calculated. As the cut-off size approaches 100% the variance in plumage scores necessarily approaches 0 and so the effect sizes will inevitably also approach 0.





Extended Data Figure 4 | Candidate models for phylogenetic confirmatory path analysis. a–n, Arrows indicate hypothesized direct links between variables.

**Extended Data Table 1 | Comparison of multivariate co-evolutionary models of male and female plumage ornamentation**

**a**

Model	$\alpha$	$\sigma$	<i>log likelihood</i>	AIC	$\Delta$ AIC
3	correlated	correlated	-39068	78153	-
2	independent	correlated	-39101	78217	64
1	independent	independent	-40140	80292	2139
5	-	correlated	-43481	86973	8820
4	-	independent	-45156	90321	12168

**b**

Parameter	Estimate
$\theta$	female = 47.52 male = 51.02
$\alpha$	$\begin{pmatrix} 2.30 & 0.90 \\ 0.90 & 1.22 \end{pmatrix}$
$\sigma$	$\begin{pmatrix} 290.55 & 216.50 \\ 216.50 & 222.17 \end{pmatrix}$

**a**, Multivariate Ornstein–Uhlenbeck (OU) models are ordered by Akaike information criterion (AIC) value from best (top) to worst. The  $\alpha$  matrix determines the strength of selection and can incorporate an interaction among traits ('correlated') or can model independent selection on each trait ('independent'). The  $\sigma$  matrix determines the amount of variation (drift) in the stochastic evolutionary process, which can be independent or correlated among traits. Brownian motion models (4 and 5) do not contain  $\alpha$  parameters. **b**, Parameter estimates for the best model (Model 3 in **a**).  $\theta$  values are the estimated trait optima towards which male and female plumage colour evolve under OU processes. The  $\alpha$  and  $\sigma$  matrices are shown for female-specific values in the top left cell, male-specific values in the bottom right cell and the interaction between trait values in the off-diagonal elements.

Extended Data Table 2 | Phylogenetic generalized least-squares models on predictors of sexual dichromatism in the Passeriformes

	Variable	N species	$\beta$	Std.Error	t-value	P
<b>a, Single predictor models</b>	Body Mass	5066	-0.76	0.17	-4.39	< 0.0001
	Wing Length	2855	-0.66	0.25	-2.66	0.011
	Degrees from equator	5831	0.60	0.09	6.77	< 0.0001
	Seasonality	5831	0.64	0.09	7.02	< 0.0001
	Clutch size	3909	0.51	0.13	4.01	< 0.0001
	Social polygyny	3504	1.51	0.15	9.83	< 0.0001
	Sexual size dimorphism	2855	0.70	0.15	4.59	< 0.0001
	Paternal care	5831	-0.99	0.12	-8.29	< 0.0001
	Cooperative Breeding	5831	-0.25	0.09	-2.77	0.007
	Migration	5830	0.81	0.09	9.37	< 0.0001
<b>b, Multiple predictor model</b>	Body size	2471	-0.95	0.25	-3.70	0.0002
	Tropical Life History	2471	-0.28	0.20	-1.40	0.16
	Sexual Selection	2471	1.76	0.19	9.20	<0.0001
	Cooperative Breeding	2471	-0.28	0.14	-2.01	0.04
	Migration	2471	0.81	0.19	4.33	<0.0001

To account for phylogenetic non-independence, models included a phylogenetic correlation structure using phylogenetic trees from <http://birdtree.org><sup>30</sup>. Provided are the model-averaged predictions of 20 models (each using a different tree), with sexual dichromatism (that is, male plumage score – female plumage score) as the dependent variable. The phylogenetic signal, estimated as Pagel's  $\lambda$  coefficient, was  $\lambda = 0.80$  (mean; range: 0.76–0.84). Predictor variables were scaled to a mean of 0 and standard deviation of 1.



**Extended Data Table 3 | Phylogenetic principal component analysis loadings and the variance explained by each component**

Predictor	Variable	Component Loadings		
		PPC1	PPC2	PPC3
Body size	Body mass	0.48	-0.13	
	Wing length	0.48	0.13	
	<b>% Variance</b>	93.15	6.85	
Tropical life history	Clutch size	-0.39	0.62	0
	Degrees from equator	-0.46	-0.27	-0.3
	Seasonality	-0.46	-0.26	0.3
	<b>% Variance</b>	76.59	17.32	6.09
Sexual selection	Sexual size dimorphism	-0.37	0.85	-0.01
	Social polygyny	-0.57	-0.27	0.51
	Paternal care	0.57	0.28	0.5
	<b>% Variance</b>	53.81	29.15	17.05

Provided are the mean values obtained from phylogenetic principal component analysis of 20 separate phylogenetic trees from <http://birdtree.org>. The first components were used in the main analysis. PPC, phylogenetic principal component.

**Extended Data Table 4 | Morphological, life-history and social correlates of plumage colour scores in passerine birds ( $N=2,471$  species)**

Predictor	Posterior mean	Lower 95% C.I.	Upper 95% C.I.	N	pMCMC
(Intercept)	45.91	40.64	51.33	987	< 0.001
sex	4.24	3.95	4.53	1018	< 0.001
body size	2.60	2.16	3.06	1023	< 0.001
life history	0.91	0.53	1.30	1024	< 0.001
sexual selection	-1.95	-2.30	-1.61	1009	< 0.001
cooperative breeding	0.40	0.13	0.68	1045	0.005
migration	-0.35	-0.72	0.01	1015	0.063
sex : body size	-1.14	-1.44	-0.86	1024	< 0.001
sex : life history	0.29	-0.09	0.67	1016	0.14
sex : sexual selection	2.70	2.41	3.00	1007	< 0.001
sex : cooperative breeding	-0.50	-0.80	-0.21	1018	0.0015
sex : migration	0.70	0.32	1.07	1015	0.0011

Shown are posterior estimates of the effect size in the full model (plus the 95% credibility intervals, the effective sample size ( $N$ ), and the probability that the effect is different from the null hypothesis). The full model was the deviance information criterion (DIC) favoured model out of all possible additive models ( $N=243$  models) that include sex as well as all possible combinations of the five predictor variables and their respective interactions with sex. Values are means of 100 separate runs on each of 100 separate phylogenetic trees from <http://birdtree.org>.

**Extended Data Table 5 | Comparison of models used in phylogenetic confirmatory path analysis**

Model	C	CICc	$\Delta$ CICc
K	28.81	67.12	0
L	27.90	68.25	1.13
N	43.04	81.35	14.23
M	42.13	82.47	15.35
J	169.00	201.23	134.11
E	184.74	214.93	147.81
H	189.26	215.41	148.29
B	207.37	235.54	168.42
D	244.17	274.37	207.25
A	277.68	303.83	236.71
F	304.60	330.75	263.63
C	377.06	401.19	334.07
I	790.13	820.32	753.20
G	1054.12	1078.24	1011.12

Models are ordered by CICc value from best (top) to worst. C indicates Fisher's C-statistic.  
 For a visual illustration of the models see Extended Data Fig. 4.



# Migratory neuronal progenitors arise from the neural plate borders in tunicates

Alberto Stolfi<sup>1</sup>, Kerrianne Ryan<sup>2</sup>, Ian A. Meinertzhagen<sup>2</sup> & Lionel Christiaen<sup>1</sup>

The neural crest is an evolutionary novelty that fostered the emergence of vertebrate anatomical innovations such as the cranium and jaws<sup>1</sup>. During embryonic development, multipotent neural crest cells are specified at the lateral borders of the neural plate before delaminating, migrating and differentiating into various cell types. In invertebrate chordates (cephalochordates and tunicates), neural plate border cells express conserved factors such as *Msx*, *Snail* and *Pax3/7* and generate melanin-containing pigment cells<sup>2–4</sup>, a derivative of the neural crest in vertebrates. However, invertebrate neural plate border cells have not been shown to generate homologues of other neural crest derivatives. Thus, proposed models of neural crest evolution postulate vertebrate-specific elaborations on an ancestral neural plate border program, through acquisition of migratory capabilities and the potential to generate several cell types<sup>5–7</sup>. Here we show that a particular neuronal cell type in the tadpole larva of the tunicate *Ciona intestinalis*, the bipolar tail neuron, shares a set of features with neural-crest-derived spinal ganglia neurons in vertebrates. Bipolar tail neuron precursors derive from caudal neural plate border cells, delaminate and migrate along the paraxial mesoderm on either side of the neural tube, eventually differentiating into afferent neurons that form synaptic contacts with both epidermal sensory cells and motor neurons. We propose that the neural plate borders of the chordate ancestor already produced migratory peripheral neurons and pigment cells, and that the neural crest evolved through the acquisition of a multipotent progenitor regulatory state upstream of multiple, pre-existing neural plate border cell differentiation programs.

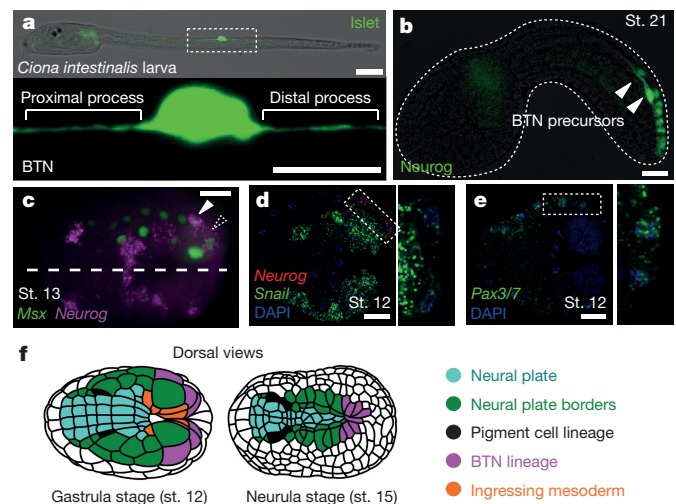
Progenitor cells that fulfil all the criteria defining the neural crest have not been observed outside vertebrates. These criteria include an embryonic origin at the lateral borders of the neural plate, epithelium-to-mesenchyme transition (EMT), migratory behaviour and the potential to differentiate into diverse cell types such as neurons, bone, cartilage and pigment cells.

In cephalochordates (amphioxus) and the tunicates *Halocynthia* and *Ciona*, a subset of neural plate border cells deploy a conserved melanocyte-specific gene network but do not migrate away from the neural tube<sup>2–4</sup>. Instead, they contribute locally to pigmented photoreceptor organs. In *Ciona*, the pigment cell precursors undergo an epithelial-to-mesenchymal transition and remain inside the neural tube lumen, but can be induced to exit the neural tube through targeted mis-expression of the mesenchyme-specific transcription factor Twist-related<sup>4</sup>. Migratory pigment cell precursors have also been reported in larvae of the tunicate *Ecteinascidia turbinata*<sup>8</sup>.

In contrast, invertebrate homologues of neural-crest-derived neurons have so far proved elusive. In tunicates, various neurons arise from the neural plate borders, but these remain in the dorsal neural tube or in the epidermis<sup>9,10</sup>, instead of delaminating and migrating as would be expected for homologues of vertebrate neural-crest-derived neurons. Migratory sensory neurons have been described in cephalochordate

embryos, but these arise from ventral epidermis, not the neural plate borders, and reinsert into the epidermis after migrating<sup>11</sup>.

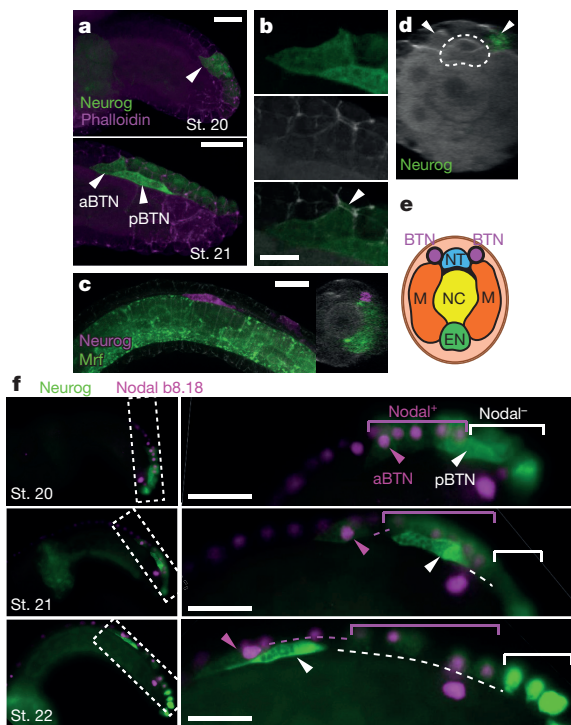
The recently identified bipolar tail neurons (BTNs)<sup>12</sup> of *Ciona* larvae form axon fascicles that extend along the length of the tail on either side of the neural tube (Fig. 1a). These neurons express the proneural basic helix–loop–helix transcription factor Neurogenin (Neurog, Fig. 1b) and the LIM-homeodomain factor Islet (Fig. 1a). Vertebrate Neurogenin and Islet orthologues are involved in specifying various neuronal subtypes including neural-crest-derived dorsal root ganglia neurons (DRGNs), which also have a bipolar or pseudo-unipolar morphology and transmit peripheral mechanosensory inputs to the central nervous system<sup>13</sup>. *Ciona* BTNs also express *Asic*, the orthologue of acid-sensing ion channels (ASICs)<sup>14</sup> that modulate touch sensitivity in vertebrate DRGNs. These parallels prompted us to investigate the embryological origins of the BTNs.



**Figure 1 | Bipolar tail neurons come from the borders of the neural plate.**

**a**, Larva with a BTN labelled by *Islet* *BTN* > *unc-76::eGFP* (green). Bottom, enlarged view of BTN above. Scale bars, 75  $\mu$ m (top); 25  $\mu$ m (bottom). **b**, Migrating BTN precursors (arrowheads) labelled by the b-line-specific *Neurog* b-line > *unc-76::Venus* reporter construct (green). Scale bar, 25  $\mu$ m. **c**, *In situ* hybridization for *Neurog* (magenta) in an embryo electroporated with *Msx* > *nls::lacZ* plasmid (immunolabelling of  $\beta$ -galactosidase in green). White arrowhead, *Msx*<sup>+</sup>/*Neurog*<sup>+</sup> BTN progenitor. Dashed arrowhead, transient *Neurog* expression in BTN progenitor's sister cell (epidermal progenitor). Dashed line, midline. Scale bar, 25  $\mu$ m. **d**, *In situ* hybridization for *Neurog* (red) and *Snail* (green). Scale bar, 25  $\mu$ m. Inset is enlarged box showing low levels of *Snail* expression in BTN progenitor. **e**, *Pax3/7* *in situ* hybridization (green). Scale bar, 25  $\mu$ m. Enlarged box inset showing *Pax3/7* expression in BTN progenitor. **f**, Adapted illustration<sup>17</sup> of embryos showing position of pigment cell and BTN progenitors (and their descendants) in the neural plate borders. Lateral views in **a**, **b**, dorsal views in **c–f**. Anterior to the left throughout; st., stage.

<sup>1</sup>Center for Developmental Genetics, Department of Biology, New York University, New York, New York 10003, USA. <sup>2</sup>Department of Psychology and Neuroscience, Life Sciences Centre, Dalhousie University, Halifax, Nova Scotia B3H 4R2, Canada.

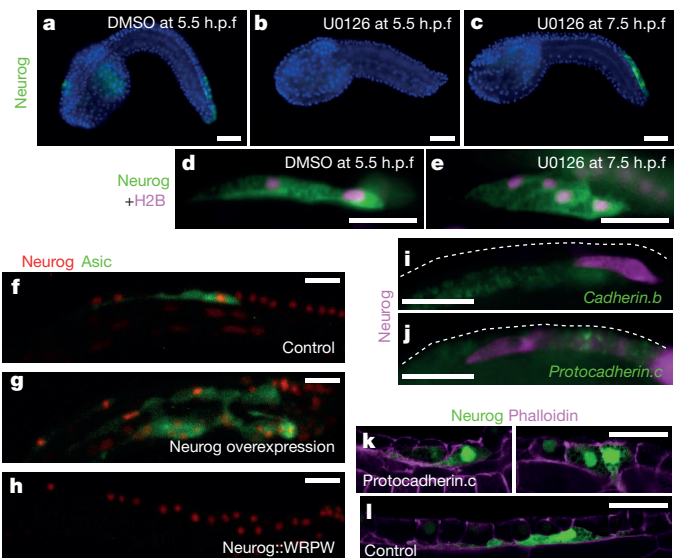


**Figure 2 | Bipolar tail neuron precursors delaminate and migrate.** **a**, Embryos electroporated with *Neurog b-line > unc-76::eGFP*. Top, BTN precursor (arrowhead) extending a lamellipodium. Bottom, BTN precursors (anterior (aBTN) and posterior (pBTN)) delaminating. Scale bar, 25  $\mu$ m. **b**, Enlarged view of aBTN in lower panel of **a**. Top, *UNC-76::eGFP*; middle, phalloidin; bottom, merged; arrowhead, part of aBTN still in the epithelium. Scale bar, 10  $\mu$ m. **c**, Embryo with paraxial mesoderm labelled by *Mrf > unc-76::eGFP* (green), BTN labelled by *Neurog b-line > unc-76::mCherry* (magenta) and phalloidin counterstain. Scale bar, 25  $\mu$ m. Right, cross-sectioned 3D image of same embryo. Only the right side of the embryo was transfected. **d**, 3D slice of embryo showing BTNs (arrowheads) outside neural tube (dotted outline). Only the right side of the embryo was transfected. **e**, Diagram of **d** showing BTNs relative to other tail tissues: neural tube (NT), notochord (NC), myoblasts (M) and endoderm (EN). **f**, Time series of different embryos co-electroporated with *Neurog b-line > unc-76::VenusYFP* (green) and *Nodal b8.18 > H2B::mCherry* (magenta). Right panels are enlarged views of the images on the left. Dashed lines indicate displacement from clonally related epidermal cells (indicated by colour-coded brackets). Scale bars, 25  $\mu$ m.

We detected the earliest expression of *Neurog* at neurulation, in the caudal-most neural/epidermal boundary cells, which express the conserved neural plate border specification genes *Msx*<sup>15</sup>, *Pax3/7* (ref. 3) and *Snail*<sup>16</sup> (Fig. 1c–f and Extended Data Fig. 1). During neurulation, these cells drive neural tube closure and their progeny eventually form the neural tube roof plate and dorsal epidermis midline<sup>17,18</sup> (Fig. 1b and Extended Data Fig. 2). BTN progenitors are thus born from the caudal extensions of the lateral borders of the neural plate (Fig. 1f).

We isolated a *Neurog cis*-regulatory element that drives reporter gene expression in this caudal neural plate border region (Extended Data Fig. 3). Using this reporter, we determined that *Neurog* expression is progressively restricted and maintained in only two cells on each side of the bilaterally symmetric embryo, born during neural tube closure (Extended Data Figs 2 and 4). We have named these the anterior (aBTN) and posterior (pBTN) BTN precursors. Shortly after the completion of neural tube closure, BTN precursors delaminate and migrate anteriorly along the paraxial mesoderm on either side of the neural tube<sup>19</sup> (Fig. 2a–f and Supplementary Videos 1–3). This is evocative of vertebrate DRGN progenitors, which migrate through paraxial mesoderm situated lateral to the neural tube.

Double-labelling with a *Nodal* reporter revealed that BTNs arise from two adjacent but clonally distinct cell lineages (Fig. 2g and Extended



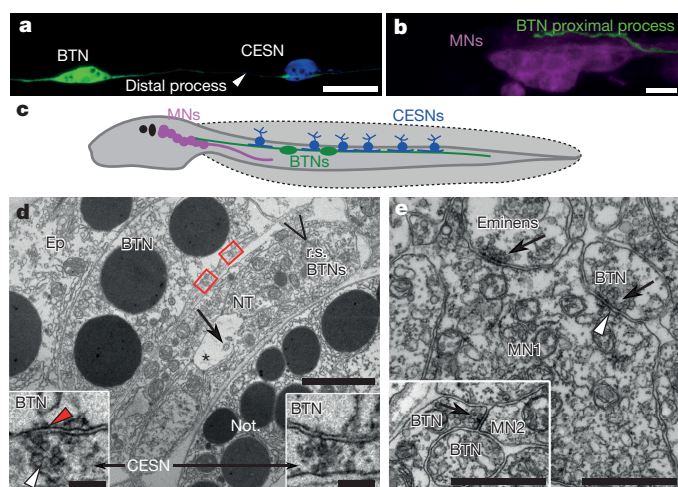
**Figure 3 | Bipolar tail neuron specification and differentiation.** **a**, Wild-type *Neurog b-line > unc-76::VenusYFP* expression (green) in embryos treated with DMSO vehicle, counterstained with DAPI (blue). **b**, *Neurog* expression was abolished in 43 of 50 embryos treated with 10  $\mu$ M MEK inhibitor U0126 at 5.5 hours post-fertilization (h.p.f.). **c**, Supernumerary BTNs were specified in 28 of 50 embryos treated with 10  $\mu$ M U0126 at 7 h.p.f. **d**, Two BTN precursors, labelled by *Neurog b-line > unc-76::VenusYFP* (green) and *Neurog b-line > H2B::mCherry* (magenta), migrating on one side of a DMSO-treated embryo. **e**, Expanded chain of four BTNs resulting from treatment with U0126 at 7 h.p.f. **f**, BTN expressing *Asic > unc-76::eGFP* reporter in embryo electroporated with *Neurog b-line > nlx::lacZ* as a control. **g**, Overexpression of *Neurog* induces specification of ectopic *Asic*<sup>+</sup> BTNs in 53 of 100 embryos. **h**, Overexpression of a dominant repressor form of *Neurog* (*Neurog::WRPW*) abolishes BTNs in 97 of 100 embryos. **i**, **j**, *In situ* hybridization reveals expression of *Cadherin.b* (green) in the neural tube but not migrating BTN precursors (**i**) and expression of *Protocadherin.c* (green) in dorsal epidermis midline but not BTNs (**j**). Embryos in **i**, **j** electroporated with *Neurog b-line > unc-76::mCherry* (immunolabelling of mCherry in magenta). **k**, Forced overexpression of *protocadherin.c* in the BTN lineage using the *Neurog b-line* driver inhibits delamination and migration of BTNs in 7 of 14 embryos. **l**, Normal BTNs as seen in 9 of 12 control embryos (overexpression of  $\beta$ -galactosidase instead). Embryos in **k**, **l** electroporated with *Neurog b-line > unc-76::VenusYFP* and *Neurog b-line > H2B::VenusYFP* (green) and counterstained with phalloidin (magenta). All scale bars 25  $\mu$ m. Embryos in **a–e**, **i**, **j** fixed at stage 22. Embryos in **f–h**, **k**, **l** at stage 23.

Data Fig. 2). The pBTN arises from the tail tip (b8.21 lineage)<sup>10</sup> and migrates to meet the b8.18-derived aBTN as it delaminates (Fig. 2a, f). Together, they continue their migration as a chain of two cells.

*Neurog* expression distinguishes the BTNs from the caudal epidermal sensory neurons (CESNs), which remain at the dorsal midline and are specified instead by an atonal homologue (*Atoh*)-dependent regulatory program<sup>10,20</sup>. We found that the onset of *Neurog* expression requires MAPK/ERK signalling (Fig. 3a, b). However, later inhibition of MAPK/ERK resulted in the upregulation of *Neurog* in non-neural cells of the lineage, converting these into supernumerary BTNs (Fig. 3c–e and Extended Data Fig. 4). In contrast, perturbing Delta/Notch signalling did not alter BTN specification or differentiation (Extended Data Fig. 5). Overexpression of *Neurog* also induced ectopic migratory *Asic*<sup>+</sup> BTN precursors (Fig. 3f, g), while BTNs were abolished through expression of a dominant repressor form of *Neurog* (*Neurog::WRPW*, Fig. 3h). In all cases, induced supernumerary BTN precursors migrated as an expanded chain of cells (Fig. 3e, g). These data indicate that sustained *Neurog* expression in caudal neural plate border cells is controlled by MAPK/ERK signalling and is necessary and sufficient for BTN specification, migration and differentiation.

In vertebrates, neural crest EMT is effected in part through differential cell adhesion, mediated by various mechanisms regulating cadherin





**Figure 4 | Synaptic connections of bipolar tail neurons.** **a**, BTN labelled by *Gad>unc-76::eGFP* (green) contacting a CESN labelled by *Slc17a6/7/8(Vglut)>unc-76::mCherry* (blue). **b**, Proximal process of BTN labelled by *Islet BTN>unc-76::mCherry* (green) contacting motor neurons (MNs) labelled by *Fg8/17/18>unc-76::eGFP* (magenta). **c**, Diagram of *Ciona* larva showing synaptic connections between CESNs in tail epidermis, BTNs and MNs. **d**, Two synaptic inputs (red boxes, insets) from the sheet-like profile of a CESN to a left-side BTN; transmission electron micrograph from wide-area montage. The profile of a second BTN axon lies out of view. Axon profiles from two right-side BTN axons (r.s. BTNs) are visible. BTNs overlie the neural tube (NT) with neural canal (marked with an asterisk) and cross-sectioned cilia (arrow). An epidermal cell (Ep) overlies the BTN. Each synapse enlarged in inset has ~52 nm diameter presynaptic vesicles (white arrowhead), and the left synapse has a postsynaptic density (red arrowhead). Scale bars, 1  $\mu$ m (inset scale bars, 0.5  $\mu$ m). Not., notochord. **e**, Synaptic input (arrow) from a BTN to the axon of a member of the most anterior pair (A11.118) of motor neurons (MN1), identified by a cumulus of ~60-nm presynaptic vesicles and a shallow postsynaptic density (arrowhead). A second input nearby originates from the axon of an eminens neuron<sup>12</sup> (arrow). Inset, synaptic input from BTN to the axon of a second pair of motor neurons (MN2). Scale bars, 1  $\mu$ m.

function<sup>21</sup>. We found that expression of *Cadherin.b*, the predominant cadherin gene expressed in the neural tube of *Ciona* embryos, is absent in BTN precursors (Fig. 3i). Moreover, BTN precursors do not express *Protocadherin.c*, a cadherin superfamily gene expressed in CESNs and epidermis midline (Fig. 3j). Overexpression of protocadherin.c protein inhibited delamination and migration of BTNs (Fig. 3k, l), suggesting that *Ciona* BTNs and vertebrate neural crest share regulatory strategies for EMT via differential cell–cell adhesion.

We observed that each BTN precursor initially migrates anteriorly with a prominent leading edge that becomes the cell's anterior neurite (or 'proximal process'), while its Golgi apparatus is located posterior to the cell nucleus. At around 12 h post-fertilization, each BTN precursor undergoes a 180° polarity inversion, with the Golgi repositioning itself anterior to the nucleus immediately before the cell begins to elaborate the posterior segment of its neurite (the 'distal process'), resulting in a bipolar morphology (Extended Data Fig. 6, Supplementary Video 4 and Supplementary Table 1). These observations suggest that a precisely timed re-orientation of cell polarity underlies the characteristic bipolar morphology of the BTNs.

At hatching, BTN cell bodies are situated in the middle of the tail along the anterior–posterior axis, with their distal processes extending towards the tail tip and proximal processes projecting towards the motor ganglion and brain (Fig. 4a–c)<sup>12</sup>. Electron microscopy confirmed that the BTN somata lie outside the neural tube and are invariably overlain by epidermal cells (Fig. 4d). BTNs lack junctions with epidermal cells and also lack cilia, thus failing to penetrate the tunic to contact the exterior. These characteristics suggest that while distal BTN neurites may be sensory, their cell bodies lack epidermal sensory receptors found in CESNs<sup>22</sup>. Along the tail, the BTNs contact overlying

CESNs, the short processes of which do not reach the motor ganglion<sup>12</sup> (Fig. 4a–c). At these contacts, synapses form from the CESN to the BTNs (Fig. 4d). Unlike the CESNs, the proximal processes of the BTNs form synaptic contacts with the motor neurons that innervate and control the tail muscles (Fig. 4b, c, e). Each BTN establishes many such contacts upon the two most anterior pairs of motor neurons, MN1 and MN2, on both the left and right sides (Fig. 4e and Extended Data Table 1). These synaptic connections are similar to those of mammalian slowly adapting type I DRGNs that, in addition to being mechanosensitive themselves, relay distinct inputs from mechanosensory Merkel cells of the epidermis<sup>23</sup>. Both tunicate CESNs and vertebrate Merkel cells arise from non-migratory epidermal cells, require *Atoh* factors for their specification and are glutamatergic in their neurotransmitter phenotype<sup>10,20,24,25</sup>. These data suggest that tunicate BTNs may thus be equivalent to vertebrate DRGNs within a homologous ascending sensory pathway (Fig. 4c).

In anamniote vertebrates, evidence for a common progenitor of intramedullary Rohon–Beard neurons (RBNs) and neural crest, in addition to other similarities between RBNs and DRGNs, indicates a deep homology between these cell types<sup>26</sup>. Fritsch and Northcutt proposed that a key step in the evolution of neural crest was the elaboration of extramedullary sensory neurons from intramedullary RBN-like neurons<sup>27</sup>. Following the Fritsch–Northcutt model, the BTNs may be derived from an 'intermediate' extramedullary neuron that evolved in the last common ancestor of Olfactores (vertebrates and tunicates) before the appearance of bona fide neural crest in the vertebrates. The migration of BTN precursors along the paraxial mesoderm, similar to later phases of DRGN migration, suggests that some of the diverse EMT and migratory behaviours displayed by vertebrate neural crest cells may pre-date the emergence of vertebrates.

Although the embryological origin (neural plate borders) and molecular signature (*Neurog*<sup>+</sup>/*Islet*<sup>+</sup>) of the BTNs of *Ciona* also support homology with RBNs, the two do in fact differ in several key aspects. First, BTNs are extramedullary neurons derived from progenitor cells that migrate along paraxial mesoderm lateral to the neural tube. Second, expression of ASICs is shared between BTNs and DRGNs, but appears absent from RBNs<sup>28</sup>. Finally, RBNs are multipolar with extensively branching peripheral neurites that innervate the overlying epidermis<sup>29</sup>, while we have not observed any peripheral neurites projecting from the bipolar/pseudounipolar BTNs.

We have revealed the developmental history of migratory neuronal progenitors that arise from the neural plate borders of tunicate embryos. Based on their embryological origin, gene expression, cell behaviour, morphology and synaptic connections, we propose that the BTNs are homologous to neural-crest-derived DRGNs. This would imply that the neural plate borders of the olfactorean ancestor gave rise to at least two types of neural crest derivatives: pigment cells and peripheral neurons (Extended Data Fig. 7).

In the invariantly developing *Ciona* embryo, the pigment cell and BTN lineages become separated early in development, but converge at a neural plate border cell identity before parting again towards distinct differentiated fates. This separation between the two lineages may represent the ancestral condition of the neural plate borders before the evolution of the neural crest in vertebrates. This would support models that propose an evolutionary origin for vertebrate neural crest through a heterochronic shift or 'intercalation' of a multipotent progenitor state downstream of neural plate border specification but upstream of cell differentiation, based on shared regulatory programs between neural crest and pluripotent cells of the early embryo<sup>1,30</sup>.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 25 March; accepted 30 September 2015.

Published online 28 October; corrected online 18 November 2015 (see full-text HTML version for details).



1. Bronner, M. E. & LeDouarin, N. M. Evolution and development of the neural crest: an overview. *Dev. Biol.* **366**, 2–9 (2012).
2. Yu, J.-K., Meulemans, D., McKeown, S. J. & Bronner-Fraser, M. Insights from the amphioxus genome on the origin of vertebrate neural crest. *Genome Res.* **18**, 1127–1132 (2008).
3. Wada, H., Holland, P. W. H., Sato, S., Yamamoto, H. & Satoh, N. Neural tube is partially dorsalized by overexpression of *HrPax-37*: the ascidian homologue of *Pax-3* and *Pax-7*. *Dev. Biol.* **187**, 240–252 (1997).
4. Abitua, P. B., Wagner, E., Navarrete, I. A. & Levine, M. Identification of a rudimentary neural crest in a non-vertebrate chordate. *Nature* **492**, 104–107 (2012).
5. Wada, H. Origin and evolution of the neural crest: a hypothetical reconstruction of its evolutionary history. *Dev. Growth Differ.* **43**, 509–520 (2001).
6. Baker, C. V. H. & Bronner-Fraser, M. The origins of the neural crest. Part II: an evolutionary perspective. *Mech. Dev.* **69**, 13–29 (1997).
7. Shimeld, S. M. & Holland, P. W. H. Vertebrate innovations. *Proc. Natl Acad. Sci. USA* **97**, 4449–4452 (2000).
8. Jeffery, W. R., Strickler, A. G. & Yamamoto, Y. Migratory neural crest-like cells form body pigmentation in a urochordate embryo. *Nature* **431**, 696–699 (2004).
9. Mazet, F. *et al.* Molecular evidence from *Ciona intestinalis* for the evolutionary origin of vertebrate sensory placodes. *Dev. Biol.* **282**, 494–508 (2005).
10. Pasini, A. *et al.* Formation of the ascidian epidermal sensory neurons: insights into the origin of the chordate peripheral nervous system. *PLoS Biol.* **4**, e225 (2006).
11. Kaltenbach, S. L., Yu, J.-K. & Holland, N. D. The origin and migration of the earliest-developing sensory neurons in the peripheral nervous system of amphioxus. *Evol. Dev.* **11**, 142–151 (2009).
12. Imai, J. H. & Meinertzhagen, I. A. Neurons of the ascidian larval nervous system in *Ciona intestinalis*: II. Peripheral nervous system. *J. Comp. Neurol.* **501**, 335–352 (2007).
13. Ma, Q., Fode, C., Guillemot, F. & Anderson, D. J. NEUROGENIN1 and NEUROGENIN2 control two distinct waves of neurogenesis in developing dorsal root ganglia. *Genes Dev.* **13**, 1717–1728 (1999).
14. Coric, T., Passamaneck, Y. J., Zhang, P., Di Gregorio, A. & Canessa, C. M. Simple chordates exhibit a proton-independent function of acid-sensing ion channels. *FASEB J.* **22**, 1914–1923 (2008).
15. Aniello, F. *et al.* Identification and developmental expression of *Ci-msxb*: a novel homologue of *Drosophila msh* gene in *Ciona intestinalis*. *Mech. Dev.* **88**, 123–126 (1999).
16. Wada, S. & Saiga, H. Cloning and embryonic expression of *Hrsna*, a snail family gene of the ascidian *Halocynthia roretzi*: implication in the origins of mechanisms for mesoderm specification and body axis formation in chordates. *Dev. Growth Differ.* **41**, 9–18 (1999).
17. Hashimoto, H., Robin, F. B., Sherrard, K. M. & Munro, E. M. Sequential contraction and exchange of apical junctions drives zipper and neural tube closure in a simple chordate. *Dev. Cell* **32**, 241–255 (2015).
18. Nicol, D. & Meinertzhagen, I. Development of the central nervous system of the larva of the ascidian, *Ciona intestinalis* L.: II. Neural plate morphogenesis and cell lineages during neurulation. *Dev. Biol.* **130**, 737–766 (1988).
19. Nakamura, M. J., Terai, J., Okubo, R., Hotta, K. & Oka, K. Three-dimensional anatomy of the *Ciona intestinalis* tailbud embryo at single-cell resolution. *Dev. Biol.* **372**, 274–284 (2012).
20. Tang, W. J., Chen, J. S. & Zeller, R. W. Transcriptional regulation of the peripheral nervous system in *Ciona intestinalis*. *Dev. Biol.* **378**, 183–193 (2013).
21. Theveneau, E. & Mayor, R. Neural crest delamination and migration: from epithelium-to-mesenchyme transition to collective cell migration. *Dev. Biol.* **366**, 34–54 (2012).
22. Torrence, S. & Cloney, R. Nervous system of ascidian larvae: caudal primary sensory neurons. *Zoomorphology* **99**, 103–115 (1982).
23. Maksimovic, S. *et al.* Epidermal Merkel cells are mechanosensory cells that tune mammalian touch receptors. *Nature* **509**, 617–621 (2014).
24. Morrison, K. M., Miesegaes, G. R., Lumpkin, E. A. & Maricich, S. M. Mammalian Merkel cells are descended from the epidermal lineage. *Dev. Biol.* **336**, 76–83 (2009).
25. Horie, T., Kusakabe, T. & Tsuda, M. Glutamatergic networks in the *Ciona intestinalis* larva. *J. Comp. Neurol.* **508**, 249–263 (2008).
26. Artinger, K. B., Chitnis, A. B., Mercola, M. & Driever, W. Zebrafish narrowminded suggests a genetic link between formation of neural crest and primary sensory neurons. *Development* **126**, 3969–3979 (1999).
27. Fritsch, B. & Northcutt, R. G. Cranial and spinal nerve organization in amphioxus and lampreys: evidence for an ancestral craniate pattern. *Acta Anat. (Basel)* **148**, 96–109 (1993).
28. Paukert, M. *et al.* A family of acid-sensing ion channels from the zebrafish: widespread expression in the central nervous system suggests a conserved role in neuronal communication. *J. Biol. Chem.* **279**, 18783–18791 (2004).
29. O'Brien, G. S. *et al.* Coordinate development of skin cells and cutaneous sensory axons in zebrafish. *J. Comp. Neurol.* **520**, 816–831 (2012).
30. Buitrago-Delgado, E., Nordin, K., Rao, A., Geary, L. & LaBonne, C. Shared regulatory programs suggest retention of blastula-stage potential in neural crest cells. *Science* **348**, 1332–1335 (2015).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** The authors would like to thank F. Razy-Krajka for assistance with Kaede photoconversion and comments on the manuscript, T. Tolkin for constructing the *Mrf* reporter plasmid, Z. Lu for ultramicrotomy, and C. Desplan, A. Di Gregorio and all members of the Christiaen and Meinertzhagen labs for feedback and suggestions. We thank H. Hashimoto, F. Robin and N. Takatori for embryo illustration template files. This work was funded by a National Science Foundation Postdoctoral Fellowship in Biology (under grant NSF-1161835) to A.S., by National Institutes of Health award GM096032 to L.C., and by grant DIS0000065 from NSERC (Ottawa) to I.A.M.

**Author Contributions** A.S., K.R., I.A.M. and L.C. designed the study, analysed the data, and wrote the paper. A.S. and K.R. performed the experiments.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to L.C. ([lc121@nyu.edu](mailto:lc121@nyu.edu)).

## METHODS

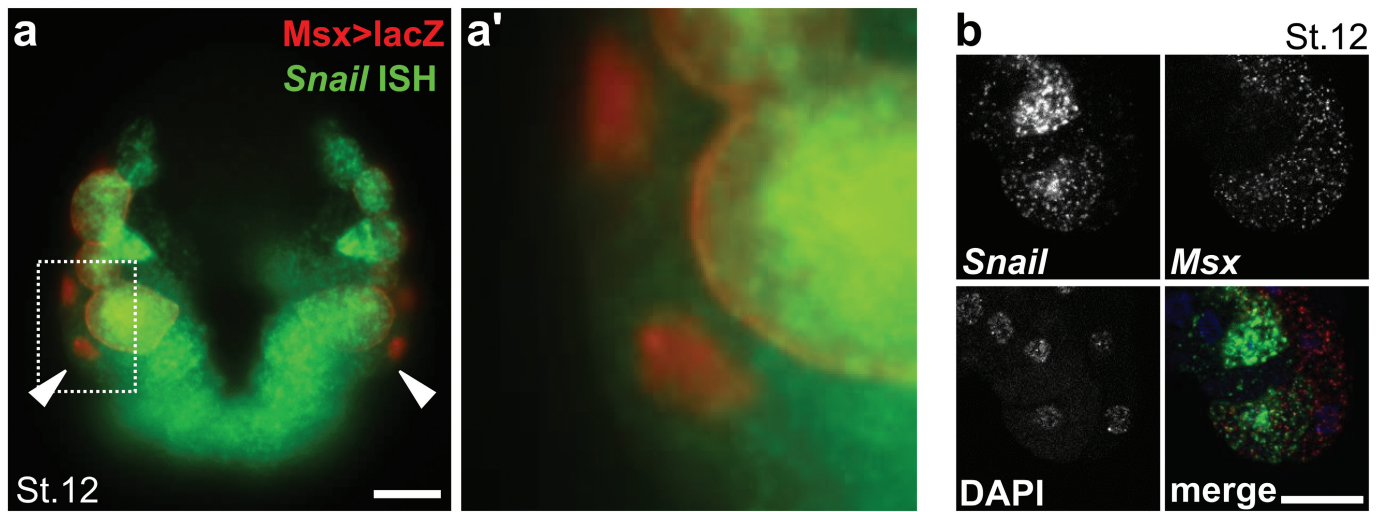
**Molecular cloning.** Reporter constructs were designed based on information of *cis*-regulatory modules (CRMs) from previously published studies on the following genes: *Islet*<sup>31</sup>, *Msx*<sup>32</sup>, *Neurog*<sup>33</sup>, *Nodal*<sup>34</sup>, *Asic*<sup>14</sup>, glutamate decarboxylase (*Gad*)<sup>35</sup>, *Slc17a6/7/8* (*Vglut*)<sup>25</sup> and *Fgf8/17/18* (ref. 36). The *Neurog b-line* CRM (Ciinte.REG.KhC6.1500090-1502346) was cloned using the following primers: *Neurog* −3,010 forward (5′-GTCTGTTTCCGCATACATGC-3′) and *Neurog* −773 reverse (5′-CTTATACGCCGAACCTCATG-3′). The *Neurog b-line* minimal CRM (Ciinte.REG.KhC6.1500090-1500501) was found to be contained within this region and cloned using *Neurog* −3,010 forward and *Neurog* −2,599 reverse (5′-GCAAAACGTTTCCCGATTTCG-3′) primers. *Neurog* CRMs were cloned upstream of the basal promoter of *Neurog* (Ciinte.REG.KhC6.1502506-1503107), cloned using the primers *Neurog* −594 forward (5′-GGTCATGCTTTGTACGTCC-3′) and *Neurog* +9 reverse (5′-ATCCAACATTTTGTAGCAAGAGC-3′), or the basal promoter of the *Zfpm* gene (also known as friend of GATA, or *Fog*)<sup>37</sup>. The full-length *Mrf* CRM (Ciinte.REG.KhC14.4311719-4314636) was cloned using the primers (5′-GCAAGCTCCTTTGGGGTTTGG-3′) and (5′-CGTATAAATATGTCAAACACTACCGGC-3′). *Caenorhabditis elegans* UNC-76 tags were fused to fluorescent proteins to ensure even labelling of axons<sup>38</sup>. Probes used for *in situ* hybridization were transcribed *in vitro* from templates obtained from previously published gene collection clones<sup>39,40</sup> for *Neurog* (R1CiGC29n04), *Pax3/7* (R1CiGC42e20), *Ebf* (R1CiGC02i14) and *Cadherin.b* (VES104\_F13) or cloned *de novo* from coding sequences for *Snail* (KH.C3.751.v1.C.SL1-1) and *Protocadherin.c* (KH.C9.32.v1.A.SL1-1). Golgi-targeting sequence was cloned from KH.C14.396.v1.B.ND1-1 cDNA (*N*-acetylglucosaminyltransferase 7, or *Galnt7*) using the primers *Galnt7* amino acid 1 forward (5′-ATGAGATTTAAAA TCGCATCAGTTTGTG-3′) and *Galnt7* amino acid 157 reverse (5′-AAGTGATATCTGTGCGTGTTCAC-3′) and fused in-frame to fluorescent proteins. *Neurog* coding sequence and *Neurog::WRPW* have been previously cloned and published<sup>41</sup>. *dnFGFR* has been previously published<sup>42</sup>, as has *Su(H)-DBM*<sup>43</sup>.

**Embryo handling, *in situ* hybridization and immunolabelling.** For purposes other than for electron microscopy (see below), eggs and embryos from wild-caught *Ciona intestinalis* (species type A, ‘robusta’) purchased from M-REP (San Diego, California) were handled according to established protocols<sup>44</sup>. Double *in situ* hybridization/immunolabelling was performed as described in previous publications<sup>45,46</sup>. Monoclonal anti-β-galactosidase (Promega catalogue number Z3781), rabbit polyclonal anti-mCherry (BioVision, accession number ACY24904), and Alexa Fluor-conjugated secondary antibodies (Life Technologies) were all used at 1:500 working dilution. Alexa Fluor-conjugated phalloidin (Life Technologies) was used at 1:50 working dilution. MEK inhibitor U0126 (Cell Signaling Technology) was resuspended as stock solution in DMSO at 10 mM concentration, and diluted to 10 μM in artificial sea water for embryo treatments. Sample sizes equal the total number of embryos present per microscope slide, unless these exceeded arbitrarily set limits of 50 or 100 embryos. No statistical methods were used to predetermine sample size and no replicates were used. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

**Fluorescence/confocal microscopy and photoconversion.** Images were captured on a Leica inverted TCS SP8 X confocal or DM2500 epifluorescence microscope. For time-lapse image capture, embryos were imaged as they developed in sea water-filled chambers on coverslip-bottom Petri dishes (MatTek). Confocal image stacks were processed in Leica Application Suite or ImageJ. Video annotations were made using Camtasia software (TechSmith). 3D slices and projections were generated using Imaris (Bitplane) or Volocity (PerkinElmer) software. Kaede::nls<sup>47</sup> was photoconverted as previously described<sup>48</sup>. Neurite lengths and Golgi apparatus positioning were measured using ImageJ. Not all cells, neurites and/or Golgi were visible in every embryo. Golgi positioning relative to BTN nuclei was measured in degrees of angle formed between a line traced anteriorly from the nucleus and another line traced through the middle of the Golgi complex. Thus, when the Golgi complex is perfectly aligned anterior to the nucleus, the angle is 0°, whereas if the Golgi complex is perfectly posterior to the nucleus, the angle is 180°. Rose plots (angle histograms) were generated in Matlab (<http://www.mathworks.com/help/matlab/ref/rose.html>).

**Electron microscopy.** Adult animals, *Ciona intestinalis* (L.), were collected by P. Darnell from Mahone Bay, Nova Scotia. Two-hour larvae reared at 18 °C in the dark were fixed at 4 °C for 1 h in 1% OsO<sub>4</sub> in 1.25% NaHCO<sub>3</sub> adjusted to pH 7.2 with HCl, followed by 2% glutaraldehyde in 0.1 M phosphate buffer. After fixation they were embedded in Epon, and a single larva cross sectioned at 60 nm in the motor ganglion and later at 100 nm down the length of the tail, and the sections post-stained for 5–6 min in freshly prepared aqueous uranyl acetate followed by 2–3 min in lead citrate. Sections were viewed using an FEI Tecnai 12 electron microscope operated at 80 kV and images captured using either a Kodak Megaview II camera using software (AnalySIS: SIS GmbH), or a Gatan 832 Orius SC1000 CCD camera using Gatan DigitalMicrograph software to compile multi-panel montages from each section. Comprehensive electron micrograph series identified the cell bodies and axons of BTNs, motor neurons and CESNs from their positions and shapes, and these in turn enabled identification of their connections (K.R. and I.A.M., manuscript in preparation).

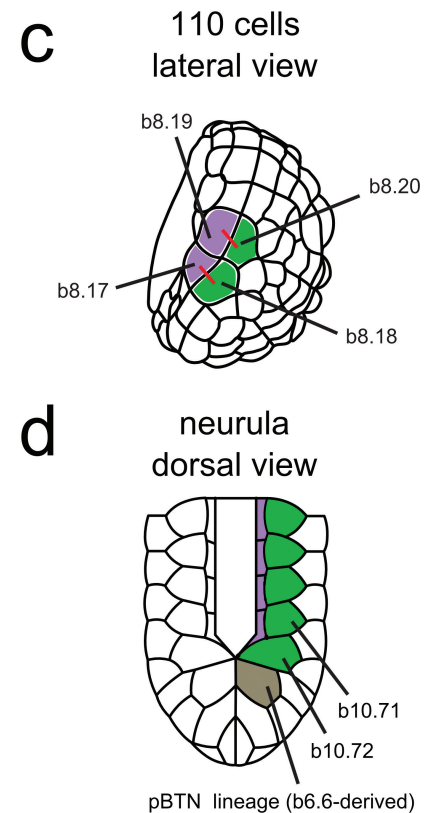
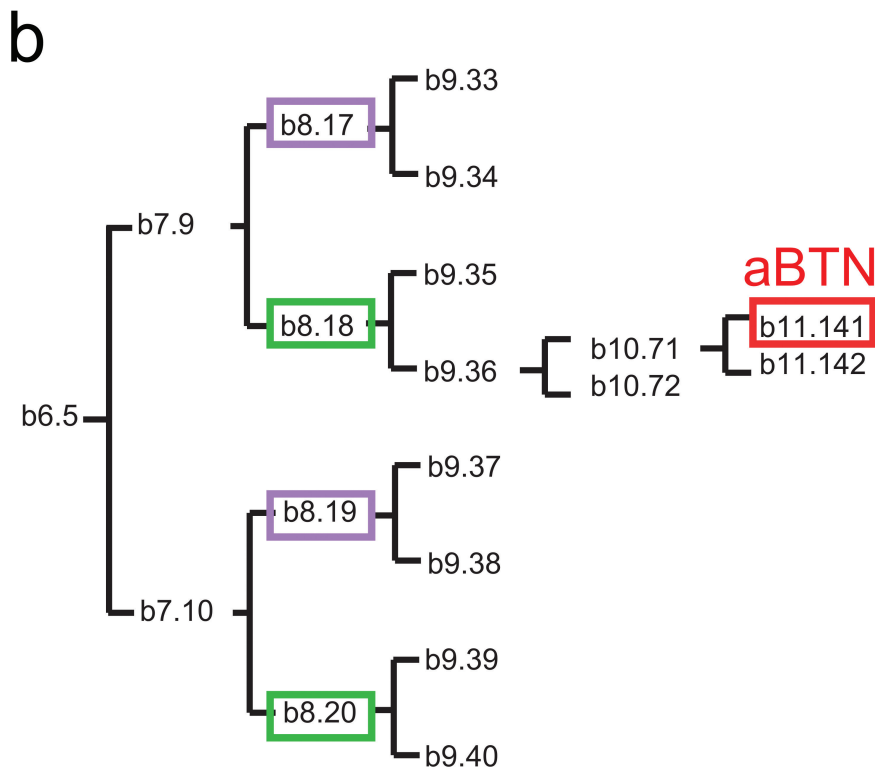
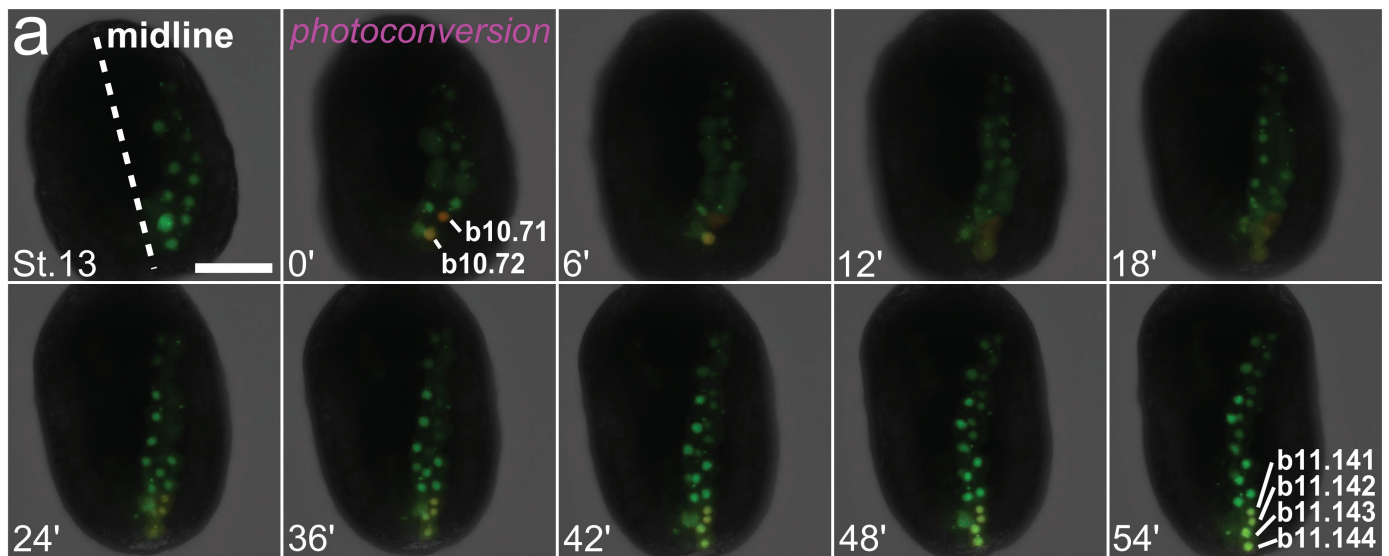
- Stolfi, A. *et al.* Early chordate origins of the vertebrate second heart field. *Science* **329**, 565–568 (2010).
- Russo, M. T. *et al.* Regulatory elements controlling *Ci-msxb* tissue-specific expression during *Ciona intestinalis* embryonic development. *Dev. Biol.* **267**, 517–528 (2004).
- Stolfi, A. & Christiaen, L. Genetic and genomic toolbox of the chordate *Ciona intestinalis*. *Genetics* **192**, 55–66 (2012).
- Khoueiry, P. *et al.* A *cis*-regulatory signature in ascidians and flies, independent of transcription factor binding sites. *Curr. Biol.* **20**, 792–802 (2010).
- Takamura, K., Minamida, N. & Okabe, S. Neural map of the larval central nervous system in the ascidian *Ciona intestinalis*. *Zool. Sci.* **27**, 191–203 (2010).
- Imai, K. S., Stolfi, A., Levine, M. & Satou, Y. Gene regulatory networks underlying the compartmentalization of the *Ciona* central nervous system. *Development* **136**, 285–293 (2009).
- Rothbacher, U., Bertrand, V., Lamy, C. & Lemaire, P. A combinatorial code of maternal GATA, Ets and β-catenin-TCF transcription factors specifies and patterns the early ascidian ectoderm. *Development* **134**, 4023–4032 (2007).
- Dynes, J. L. & Ngai, J. Pathfinding of olfactory neuron axons to stereotyped glomerular targets revealed by dynamic imaging in living zebrafish embryos. *Neuron* **20**, 1081–1091 (1998).
- Satou, Y. *et al.* A cDNA resource from the basal chordate *Ciona intestinalis*. *Genesis* **33**, 153–154 (2002).
- Roure, A. *et al.* A multicassette Gateway vector set for high throughput and comparative analyses in *Ciona* and vertebrate embryos. *PLoS ONE* **2**, e916 (2007).
- Stolfi, A., Wagner, E., Taliaferro, J. M., Chou, S. & Levine, M. Neural tube patterning by Ephrin, FGF and Notch signaling relays. *Development* **138**, 5429–5439 (2011).
- Davidson, B., Shi, W., Beh, J., Christiaen, L. & Levine, M. FGF signaling delineates the cardiac progenitor field in the simple chordate, *Ciona intestinalis*. *Genes Dev.* **20**, 2728–2738 (2006).
- Hudson, C. & Yasuo, H. A signalling relay involving Nodal and Delta ligands acts during secondary notochord induction in *Ciona* embryos. *Development* **133**, 2855–2864 (2006).
- Christiaen, L., Wagner, E., Shi, W. & Levine, M. The sea squirt *Ciona intestinalis*. *Cold Spring Harb. Protoc.* **2009**, pdb.emo138 (2009).
- Beh, J., Shi, W., Levine, M., Davidson, B. & Christiaen, L. *FoxF* is essential for FGF-induced migration of heart progenitor cells in the ascidian *Ciona intestinalis*. *Development* **134**, 3297–3305 (2007).
- Ikuta, T. & Saiga, H. Dynamic change in the expression of developmental genes in the ascidian central nervous system: revisit to the tripartite model and the origin of the midbrain–hindbrain boundary region. *Dev. Biol.* **312**, 631–643 (2007).
- Ando, R., Hama, H., Yamamoto-Hino, M., Mizuno, H. & Miyawaki, A. An optical marker based on the UV-induced green-to-red photoconversion of a fluorescent protein. *Proc. Natl Acad. Sci. USA* **99**, 12651–12656 (2002).
- Razy-Krajka, F. *et al.* Collier/OLF/EBF-dependent transcriptional dynamics control pharyngeal muscle specification from primed cardiopharyngeal progenitors. *Dev. Cell* **29**, 263–276 (2014).
- Nishida, H. Cell division pattern during gastrulation of the ascidian, *Halocynthia roretzi*. *Dev. Growth Differ.* **28**, 191–201 (1986).
- Bone, Q. The central nervous system in amphioxus. *J. Comp. Neurol.* **115**, 27–64 (1960).



**Extended Data Figure 1 | *In situ* hybridization of neural plate border markers *Snail* and *Msx*.** **a**, Immunolabelling for  $\beta$ -galactosidase (red) and *in situ* hybridization for *Snail* mRNA (green) in stage 12 embryo electroporated with *Msx>lacZ*, revealing *Snail* expression in the BTN progenitors (b9.36 cells, arrowheads). Dashed area enlarged in **a'**.

**b**, Double *in situ* hybridization for *Snail* (green on merged image) and *Msx* (red on merged image) in stage 12 embryos counterstained with DAPI (blue on merged image), showing co-expression in neural plate border cells, including BTN progenitors. Scale bars, 25  $\mu$ m.





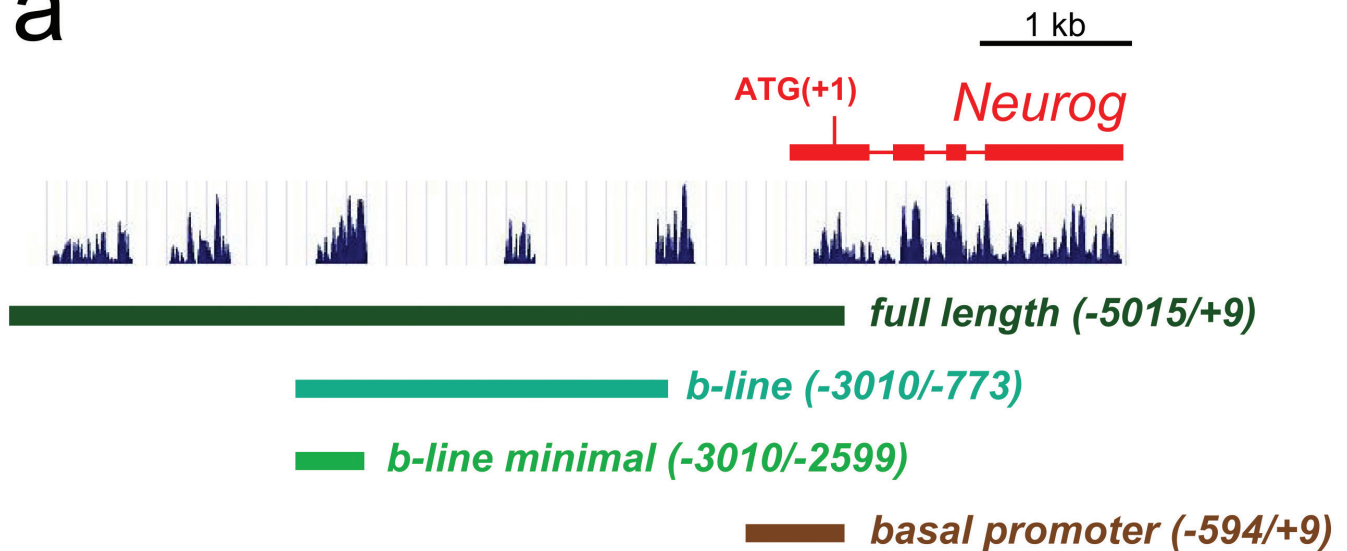
32 cells	64 cells	110 cells	gastrula	neurula	tailbud
4 hpf	4 hpf	4 hpf	5 hpf	6 hpf	7 hpf

#### Extended Data Figure 2 | Lineage tracing of b9.36 descendants.

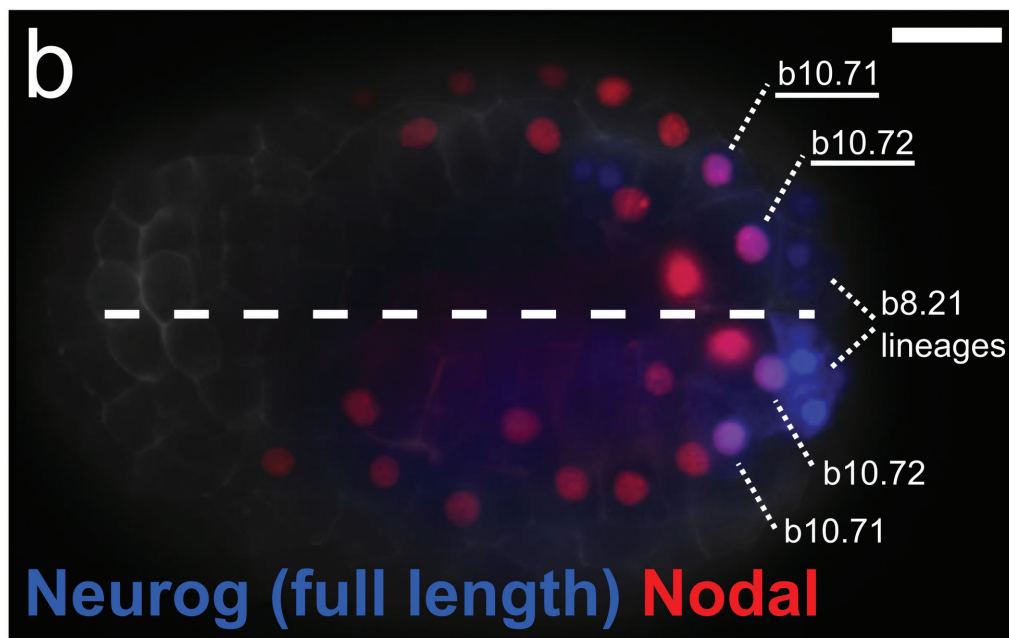
**a**, Photoconversion of Kaede::nls driven by the *Msx* driver was used to follow the cell divisions of the BTN progenitors from the late gastrula stage to the early tailbud stage. Both b10.71 and b10.72 divide once. b11.141 will give rise to a definitive anterior BTN (see Extended Data Fig. 4). Numbers in each panel represent time in minutes elapsed from the initial photoconversion event. Scale bar, 50  $\mu$ m. **b**, Lineage tree showing specification of aBTNs in relation to other cells of the posterior neural plate borders. For simplicity,

only one side of the embryo is depicted. **c**, Lateral view of a 110-cell-stage embryo showing the positions of blastomeres in **b**. Red lines connect sibling cells. **d**, Dorsal view of a neurula-stage embryo showing zipper of posterior neural-plate-border-derived capstone cells<sup>18</sup> as neural tube closure is initiated. Panels **b** and **d** are courtesy of H. Hashimoto and F. Robin (University of Chicago) and N. Takatori (Tokyo Metropolitan University), and partially modelled after ref. 17. Panel **c** modelled after ref. 49.

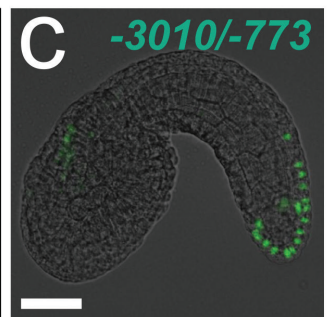
a



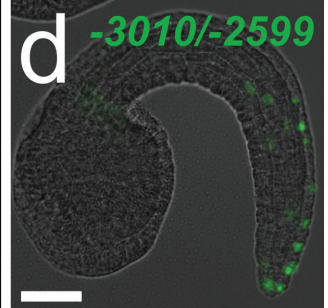
b



c

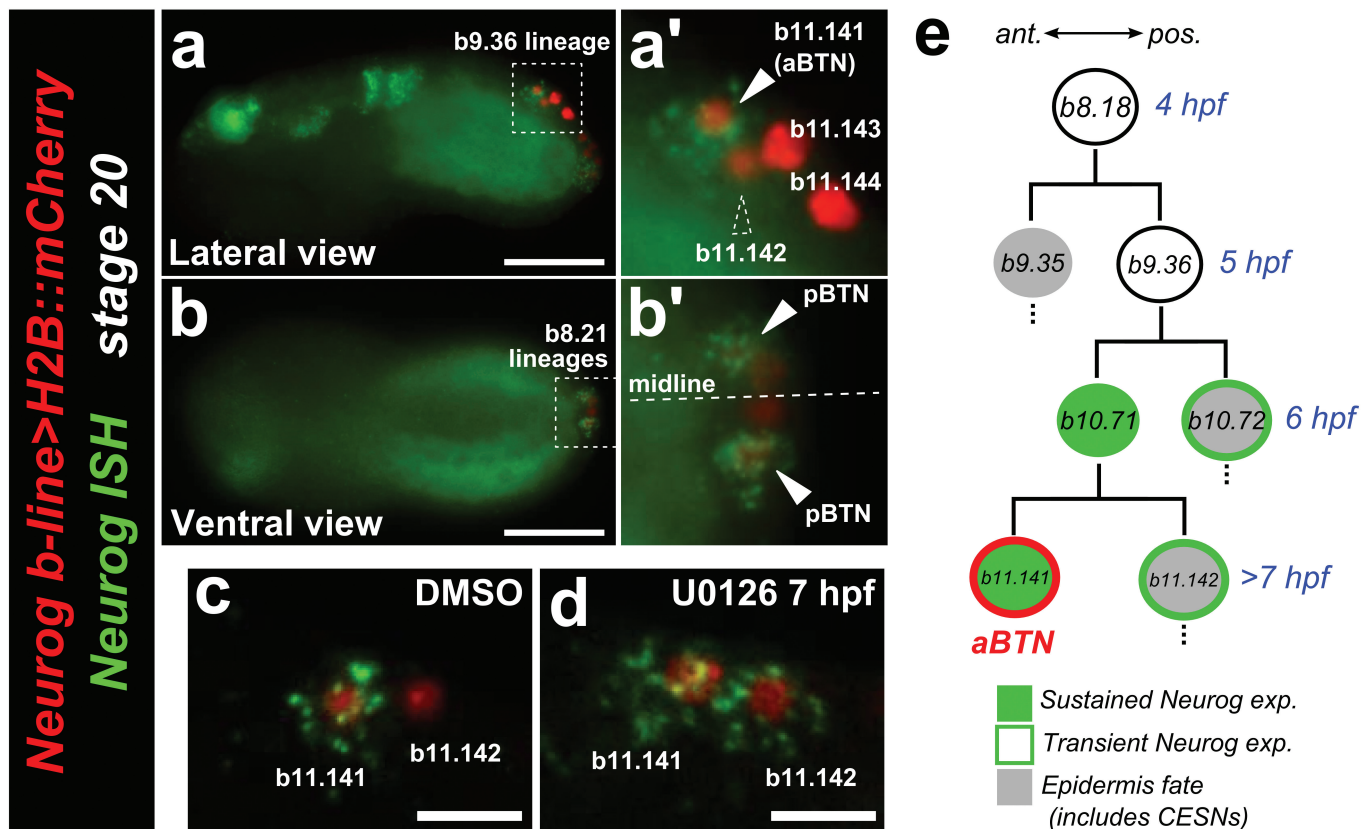


d



**Extended Data Figure 3 | *Neurog* cis-regulatory sequences.** **a**, Schematic diagram representing *Neurog* locus and 5' *cis*-regulatory sequences including *b*-line and *b*-line minimal *cis*-regulatory modules. Peaks represent nucleotide sequence conservation with *Ciona savignyi* genome. **b**, Late gastrula embryo (stage 13) electroporated with full-length *Neurog* (blue) and *Nodal*

*b*-line (red) reporter constructs. Reporter co-expression is seen in b9.36 descendants on either side of the neural plate. *Neurog* expression also marks tail-tip lineages of uncertain provenance, previously reported to be descended from b8.21 (ref. 10). Scale bar, 25 μm. **c**, *Neurog b*-line reporter. **d**, *Neurog b*-line minimal reporter. Scale bars in **c**, **d**, 50 μm.

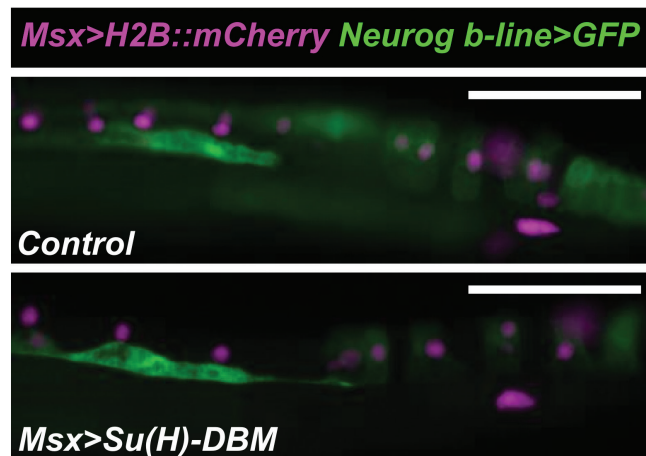


**Extended Data Figure 4 | Spatiotemporal restriction of *Neurog* expression.**

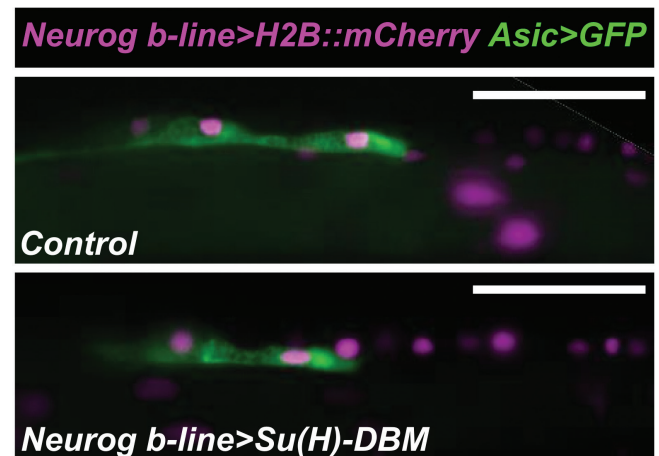
**a**, Lateral view of *in situ* hybridization (ISH) for *Neurog* (green) in embryo electroporated with *Neurog b-line>H2B::mCherry* (red) shows that *Neurog* expression is selectively maintained in only a subset of initially *Neurog*-expressing neural plate border cells. **a'**, In the b9.36 lineage, the anterior-most cell (b11.141, solid arrowhead) is always the sole one to express *Neurog* at this stage, and will go on to become the anterior BTN. Dashed arrowhead indicates b11.142, the sister cell of b11.141, which has downregulated *Neurog* relative to its sibling. **b**, **b'**, The identities of the cells in the tail tip (presumed b8.21-derived) lineages are unclear, but *Neurog* is similarly

restricted (arrowheads) to a single cell on either side of the midline, which we interpret as the definitive posterior BTNs. **c**, Control embryo treated with DMSO vehicle, showing wild-type pattern of *Neurog* expression only in b11.141. **d**, *Neurog* is expanded to b11.142 upon treatment with the MEK inhibitor U0126 at 7 h.p.f. This condition also results in specification of supernumerary BTNs, presumably due to expanded *Neurog* expression (see text for details). Thus, downregulation of *Neurog* in b11.142 also requires MEK/ERK signalling. **e**, Diagram of the aBTN lineage, descended from the b8.18 blastomere. Scale bars in **a**, **b**, 25  $\mu$ m. Scale bars in **c**, **d**, 10  $\mu$ m.

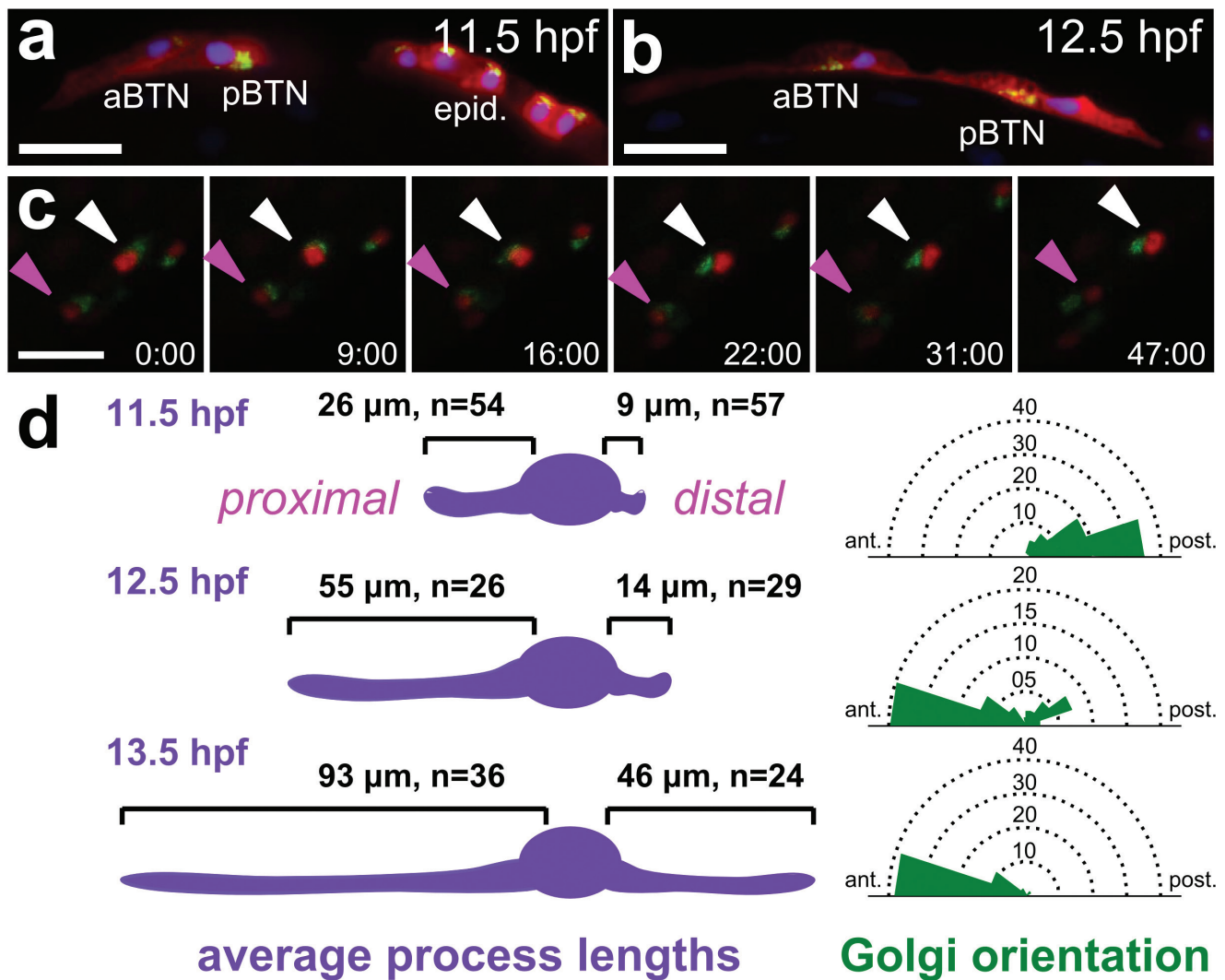


**a**

**Extended Data Figure 5 | Perturbation of Notch signalling does not alter *Neurogenin* expression or bipolar tail neuron specification and differentiation.** **a**, Top, lateral view of a stage 23 embryo electroporated with *Msx>H2B::mCherry* (magenta nuclei), *Neurog b-line>unc-76::eGFP* (green) and *Msx>nls::lacZ*, serving as the wild-type control condition. Bottom, embryo electroporated with same reporters as upper panel, plus *Msx>Su(H)-DBM*, which encodes a DNA-binding mutant form of the Notch

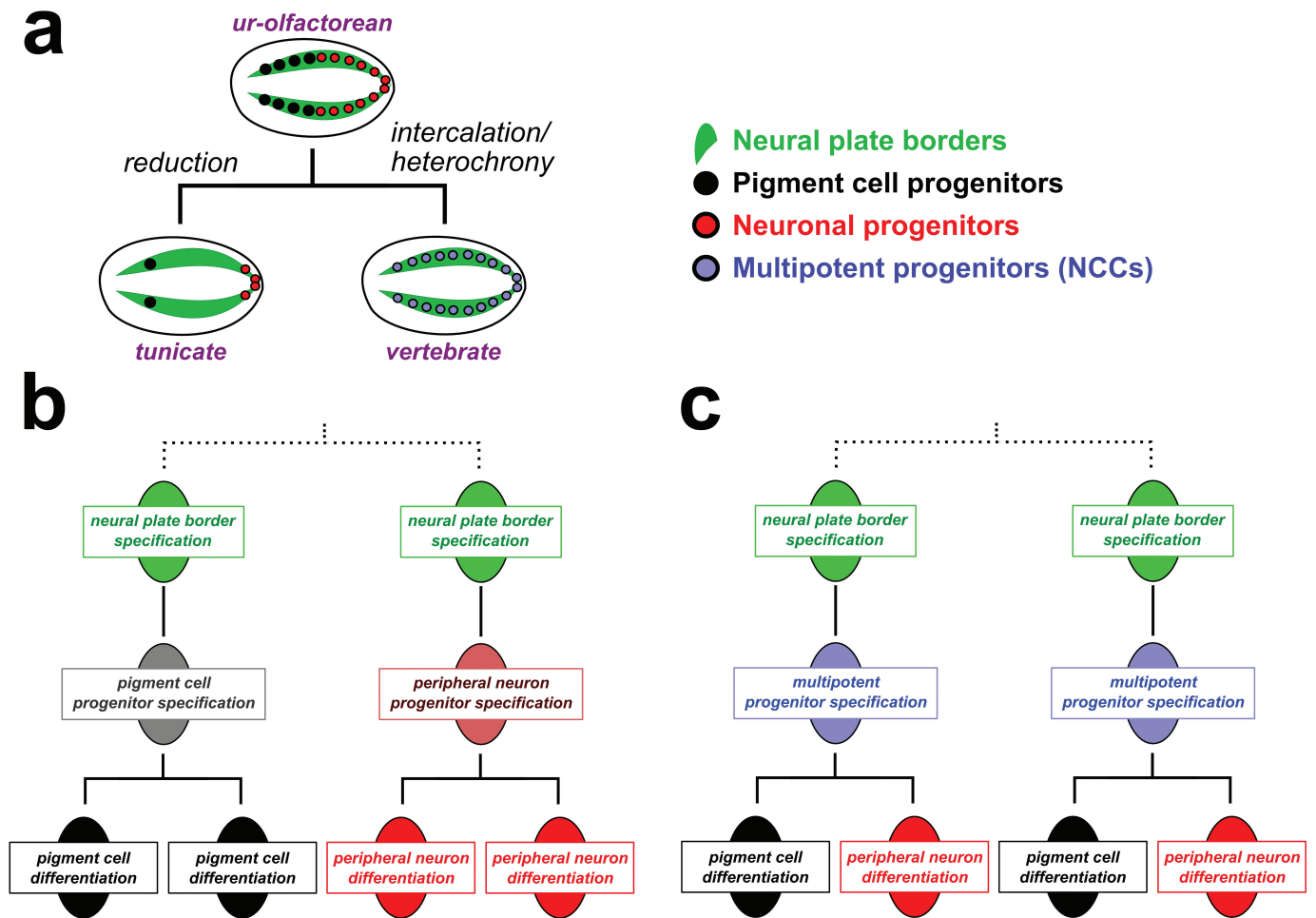
**b**

co-activator Rbpj. No discernable difference in *Neurog* activation or BTN specification was observed between control and *Su(H)-DBM* conditions (1 of 32 versus 2 of 42 embryos showing ectopic *Neurog*<sup>+</sup> BTNs, respectively). **b**, Late overexpression of *Su(H)-DBM* using the *Neurog b-line* driver similarly did not alter BTN specification/differentiation, as monitored by *Asic>unc-76::eGFP* reporter expression (0 of 50 control versus 0 of 50 *Su(H)-DBM* embryos showed ectopic *Asic*<sup>+</sup> BTNs). Scale bars, 50  $\mu$ m.



**Extended Data Figure 6 | Cell polarity and morphogenesis of bipolar tail neurons.** **a**, Embryo at 11.5 h.p.f. (18°C) with BTNs displaced from clonally related epidermal cells (epid.) labelled by UNC-76::VenusYFP (red), Galnt7 $\Delta$ C::CFP (green), and H2B::mCherry (blue) driven by *Neurog b-line* cis-regulatory module. Targeted localization of CFP by the Galnt7 N-terminal signal sequence reveals polarized subcellular distribution of Golgi apparatus on posterior side of BTN nuclei as migration and proximal process extend in an anterior direction. This is distinct from the apical (dorsal) location of the Golgi apparatus in epidermal cells. **b**, Embryo at 12.5 h.p.f. (18°C) showing 180° inversion of Golgi apparatus localization to the anterior side of the nucleus, immediately preceding distal process extension. Scale bars in **a**, **b**, 50  $\mu\text{m}$ . **c**, Still frames from a confocal image stack time lapse movie (Supplementary Video 4) showing inversion of Golgi

complex (Galnt7 $\Delta$ C::VenusYFP, green) relative to nuclei (H2B::mCherry, red) in migrating BTNs. Time lapse imaging initiated at 11.5 h.p.f. (18°C). Time in minutes elapsed from start shown at bottom right of each panel. Anterior BTN (aBTN) indicated by magenta arrowhead, posterior BTN (pBTN) indicated by white arrowhead. Scale bar, 25  $\mu\text{m}$ . **d**, Diagram showing correlation of average length of proximal (left) and distal (right) processes and angle of Golgi apparatus location relative to cell nucleus along the anterior–posterior axis in BTNs at different time points. Locations of Golgi apparatus represented by rose plots of bins of 20° spanning anterior (0°) and posterior (180°) endpoints around dorsal edge of BTN nucleus. Bin diameters indicate number of cells. Embryos analysed belong to the same pool as embryos in **a** and **b**. See Supplementary Table 1 for source data.



**Extended Data Figure 7 | Proposed evolution of neural crest through the acquisition of multipotency by neural plate border cells.** **a**, Cartoon diagram depicting a hypothetical path for neural plate border and neural crest evolution, starting with the reconstructed last common olfactorean ancestor, which could have had neural plate borders lined with committed progenitor cells giving rise to several pigmented ocelli and BTN-like peripheral neurons, a condition that may be conserved in extant cephalochordates<sup>50</sup>. These cells would have been reduced in the highly miniaturized embryos of extant tunicates, while vertebrates are proposed to have co-opted a mesenchymal, multipotency program to bestow these cells with the potential to give rise to pigment cells, peripheral neurons or other derivatives, after a prolonged

period of EMT and migration. **b**, Diagram representing idealized cell lineages in the neural plate borders of tunicate and hypothetical urolfactorean ancestor, in which segregated lineages at the neural plate borders give rise to committed pigment cell or peripheral neuronal progenitors. **c**, Diagram of simplified neural crest cell lineage deploying a multipotency program downstream of neural plate border specification and upstream of cell differentiation. Thus, neural crest cells could have evolved through redeployment of a multipotency program (intercalation hypothesis)<sup>1</sup>, or through its maintenance from earlier embryonic stages (heterochrony hypothesis)<sup>30</sup>.



Extended Data Table 1 | Synaptic input from bipolar tail neurons to motor neurons, identified by electron microscopy

Postsynaptic motor neuron identity	Synapse partnership	Number of synapses	Total number of sections with synaptic profile
MN1 Left (A11.118)	BTN1-->MN1L	27	134
	BTN3-->MN1L	21	88
	<b>Total</b>	<b>48</b>	<b>222</b>
MN1 Right (A11.118)	BTN1-->MN1R	3	14
	BTN2-->MN1R	22	94
	BTN3-->MN1R	1	4
	BTN4-->MN1R	11	55
	<b>Total</b>	<b>37</b>	<b>167</b>
MN2 Left (A10.57)	BTN1-->MN2L	10	51
	BTN3-->MN2L	6	30
	<b>Total</b>	<b>16</b>	<b>81</b>
MN2 Right (A10.57)	BTN2-->MN2R	17	90
	BTN4-->MN2R	10	73
	<b>Total</b>	<b>27</b>	<b>163</b>
MN3 Left	BTN1-->MN3L	1	2
	<b>Total</b>	<b>1</b>	<b>2</b>
MN4 Left	BTN1-->MN4L	2	9
	<b>Total</b>	<b>2</b>	<b>9</b>
MN4 Right	BTN2-->MN4R	2	5
	BTN4-->MN4R	1	2
	<b>Total</b>	<b>3</b>	<b>7</b>
MN5 Left	BTN1-->MN5L	1	3
	<b>Total</b>	<b>1</b>	<b>3</b>
MN5 Right	BTN4-->MN5R	1	3
	<b>Total</b>	<b>1</b>	<b>3</b>

BTN, bipolar tail neuron. MN, motor neuron. Axons of BTN1 and BTN3 lie on the left hand side of the embryo, and BTN2 and BTN4 on the right. The axons are not traced to their somata to indicate which would be anterior and posterior.

# Decapentaplegic and growth control in the developing *Drosophila* wing

Takuya Akiyama<sup>1</sup> & Matthew C. Gibson<sup>1,2</sup>

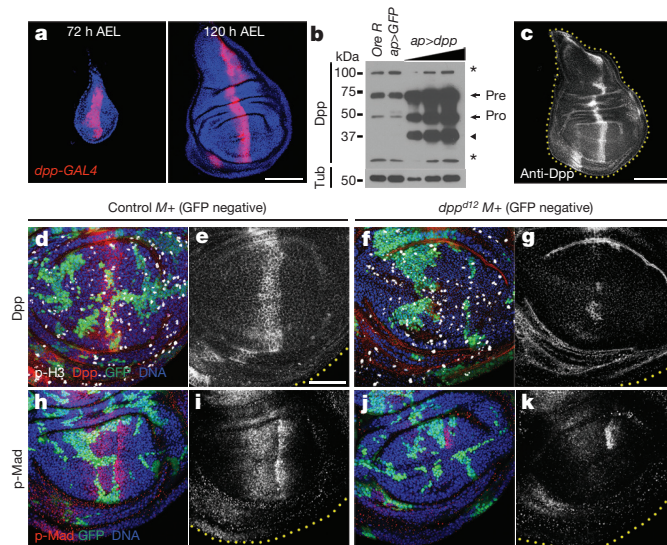
As a central model for morphogen action during animal development, the bone morphogenetic protein 2/4 (BMP2/4)-like ligand Decapentaplegic (Dpp) is proposed to form a long-range signalling gradient that directs both growth and pattern formation during *Drosophila* wing disc development<sup>1–6</sup>. While the patterning role of Dpp secreted from a stripe of cells along the anterior–posterior compartmental boundary is well established<sup>1,2,6</sup>, the mechanism by which a Dpp gradient directs uniform cell proliferation remains controversial and poorly understood<sup>7–13</sup>. Here, to determine the precise spatiotemporal requirements for Dpp during wing disc development, we use CRISPR–Cas9-mediated genome editing to generate a flippase recognition target (*FRT*)-dependent conditional null allele. By genetically removing Dpp from its endogenous stripe domain, we confirm the requirement of Dpp for the activation of a downstream phospho-Mothers against dpp (p-Mad) gradient and the regulation of the patterning targets *spalt* (*sal*), *optomotor blind* (*omb*; also known as *bifid*) and *brinker* (*brk*). Surprisingly, however, third-instar wing blade primordia devoid of compartmental *dpp* expression maintain relatively normal rates of cell proliferation and exhibit only mild defects in growth. These results indicate that during the latter half of larval development, the Dpp morphogen gradient emanating from the anterior–posterior compartment boundary is not directly required for wing disc growth.

Morphogens, signalling molecules secreted from a localized source to form gradients of activity, are proposed to coordinately control growth and patterning in diverse organismal systems<sup>1,2,6,14–17</sup>. In *Drosophila melanogaster*, the BMP2/4-like ligand Dpp is highly expressed in a row of cells at the anterior–posterior (A/P) compartment border in third-instar wing discs (72–120 h after egg laying (AEL); Fig. 1a). The secreted ligand is proposed to emanate from this position to form a long-range gradient that directs uniform growth and concentration-dependent patterning<sup>1–6</sup>. Although the requirements for Dpp in disc patterning are widely accepted<sup>1,2,6</sup>, precisely how a gradient of Dpp might direct homogenous cell proliferation is controversial<sup>7–13</sup>. The general requirements for Dpp in growth are clear; imaginal discs fail to develop in *dpp* mutant larvae<sup>18</sup> and ectopic expression of Dpp is sufficient to trigger overgrowth in lateral regions of the wing disc<sup>19</sup>. Nevertheless, owing to a lack of methods for the detection and disruption of Dpp, the mechanism by which its downstream activity gradient directs uniform cell proliferation remains unknown.

To visualize Dpp directly, we first generated a polyclonal antibody (anti-Dpp) that recognizes the Dpp prodomain on western blots and labels the expected compartmental stripe domain in fixed tissues (Fig. 1b, c and Extended Data Fig. 1). Next, to validate the specificity of anti-Dpp, we induced mitotic cell clones homozygous for the hypomorphic allele *dpp*<sup>d12</sup> (ref. 20). As expected, *dpp*<sup>d12</sup> mutant clones that impinged on the stripe domain correlated with a pronounced reduction of Dpp and p-Mad levels, indicating that both Dpp expression and the activity gradient downstream of Dpp signalling were abolished

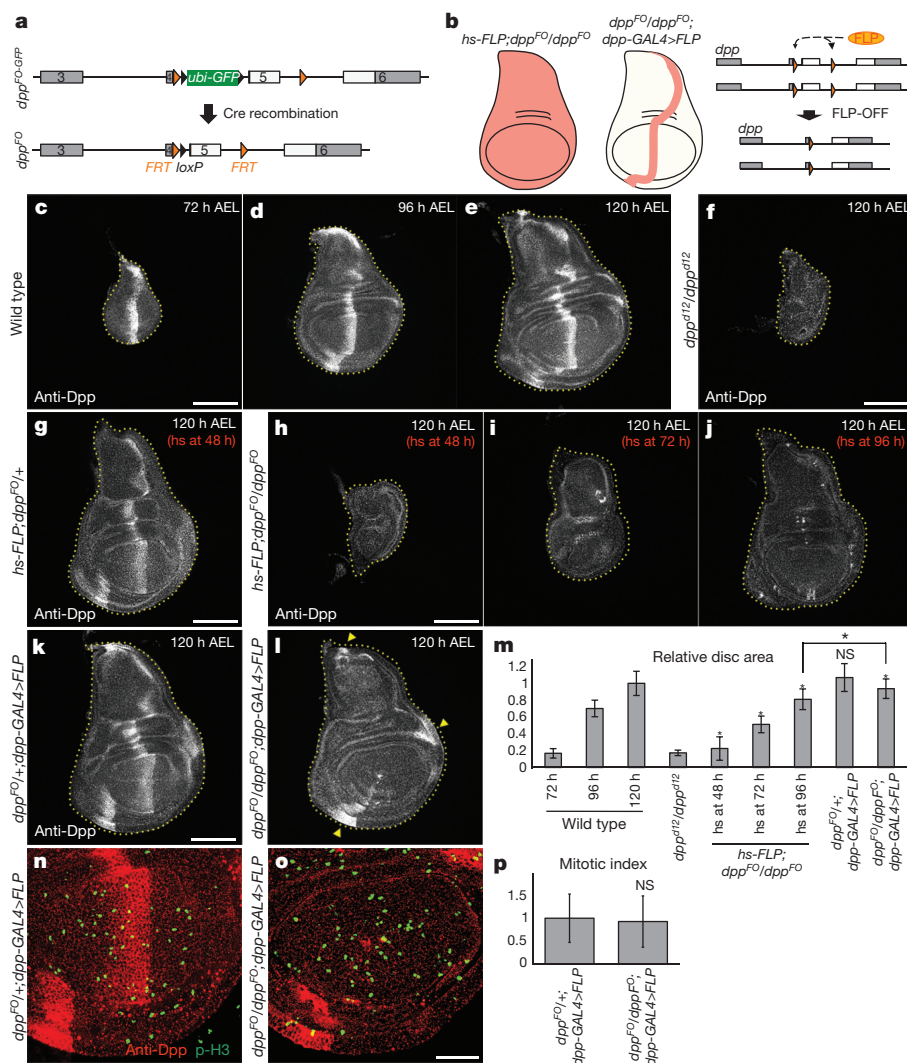
(Dpp *n* = 49, p-Mad *n* = 32; Extended Data Fig. 2). Surprisingly, however, loss of *dpp* at the compartmental boundary had minimal effects on growth and cell proliferation in third-instar discs (*n* = 80; Extended Data Fig. 2). We extended these results by generating larger *dpp*<sup>d12</sup> mutant clones using the *Minute* technique (*M*(2)25A; Fig. 1d–k)<sup>21,22</sup>. Consistent with our conventional clonal analysis, extensive loss of the Dpp stripe (*n* = 24) and its associated p-Mad activity gradient (*n* = 20) did not strongly affect proliferation in adjacent cell populations (*n* = 44; Fig. 1d–k). Indeed, only when clones were induced very early in development did we observe severely reduced wing discs.

The results described earlier indicate that the compartmental stripe of Dpp expression is not essential for growth in third-instar wing discs. However, *dpp*<sup>d12</sup> mutant clone analysis entails some limitations, including the hypomorphic nature of the lesion and the lack of precise spatiotemporal control over clone induction. Thus, to eliminate Dpp expression more precisely from the entire stripe domain, we used CRISPR–Cas9-mediated homologous recombination to generate a conditional null allele that harbours *FRT* sequences flanking the first *dpp* coding exon (*dpp*<sup>FLP-OFF</sup> (*dpp*<sup>FO</sup>); Fig. 2a, Extended Data Fig. 3a and



**Figure 1 | *dpp*<sup>d12</sup> mutant clones have little effect on wing disc growth.** **a**, Wing discs grow dramatically during the third larval instar (72–120 h AEL). *dpp*-expressing cells are visualized with *dpp*-GAL4 > UAS-GFP (red). Scale bar, 100 μm. **b**, Anti-Dpp recognizes the Dpp precursor (Pre) and prodomain (Pro) in western blots. An arrowhead indicates a previously unidentified Dpp product. Asterisks indicate non-specific bands. α-Tubulin (Tub): loading control. **c**, Dpp expression in wild-type wing disc. Scale bar, 100 μm. **d–k**, *dpp*<sup>d12</sup> mutant clones. Control (**d**, **e**, **h**, **i**) and *dpp*<sup>d12</sup> mutant (**f**, **g**, **j**, **k**) clones were stained with anti-Dpp (**d–g**) and anti-p-Mad (**h–k**). Phospho-histone 3 (p-H3; white) labels mitotic cells (**d**, **f**). Scale bars, 50 μm. Anterior is to the left in all figures.

<sup>1</sup>Stowers Institute for Medical Research, Kansas City, Missouri 64110, USA. <sup>2</sup>Department of Anatomy and Cell Biology, The University of Kansas School of Medicine, Kansas City, Kansas 66160, USA.



**Figure 2 | Eliminating the Dpp stripe causes only mild growth defects.** **a**, Design of  $dpp^{FO}$ . **b**, General strategy for either global or stripe-specific disruption of  $dpp$  through controlled expression of the FLP recombinase. **c–e**, Anti-Dpp staining of wing discs during the third instar. **f**, Anti-Dpp staining is lost in  $dpp^{d12}/dpp^{d12}$  mutant wing discs. **g**, Control wing disc from  $hs-FLP; dpp^{FO}/+$  larvae heat shocked at 48 h AEL. **h–j**,  $hs-FLP; dpp^{FO}/dpp^{FO}$  wing discs show varying degrees of size reduction after heat shock at the time points indicated. **k–l**, Dpp stripe expression is maintained in controls (**k**) but eliminated from the wing blade region in  $dpp^{FO}/dpp^{FO}$ ;  $dpp-GAL4/UAS-FLP$  wing discs (**l**). Scale bars, 100  $\mu$ m. **m**, Size comparison between wing discs of the genotypes indicated. Mean  $\pm$  standard deviation (s.d.). \* $P < 0.001$ , not significant (NS), two-sided Student's  $t$ -test. **n–o**, Mitotic cells in  $dpp^{FO}/+$ ;  $dpp-GAL4/UAS-FLP$  controls (**n**) and  $dpp^{FO}/dpp^{FO}$ ;  $dpp-GAL4/UAS-FLP$  wing discs (**o**) wing discs labelled with anti-p-H3 (green) and anti-Dpp (red). Scale bar, 50  $\mu$ m. **p**, Mitotic index in  $dpp^{FO}/+$ ;  $dpp-GAL4/UAS-FLP$  controls ( $n = 57$ ) and  $dpp^{FO}/dpp^{FO}$ ;  $dpp-GAL4/UAS-FLP$  wing discs ( $n = 54$ ). Scale bar, 100  $\mu$ m. \* $P < 0.001$ , not significant (NS), two-sided Student's  $t$ -test.

Methods). The  $dpp^{FO}$  allele is homozygous viable, and thus  $dpp$ -null mutant cells can be generated in precise spatial and temporal patterns by using controlled expression of the FLP recombinase to direct excision of the genomic region between the  $FRT$  sites (Fig. 2b and Extended Data Figs 3b, 4a–d). Similar to previously characterized  $dpp$  mutants,  $hs-FLP; dpp^{FO}/dpp^{FO}$  larvae exhibited severely reduced wing discs and loss of anti-Dpp staining when subjected to heat-shock-induced disruption of  $dpp$  during early larval development at 48 h AEL (Fig. 2c–h, m). We observed a similar loss of Dpp but less pronounced growth defects after Dpp removal at later time points (72 and 96 h AEL), indicating that there is a continuous requirement for  $dpp$  expression during wing disc development (Fig. 2i–j, m).

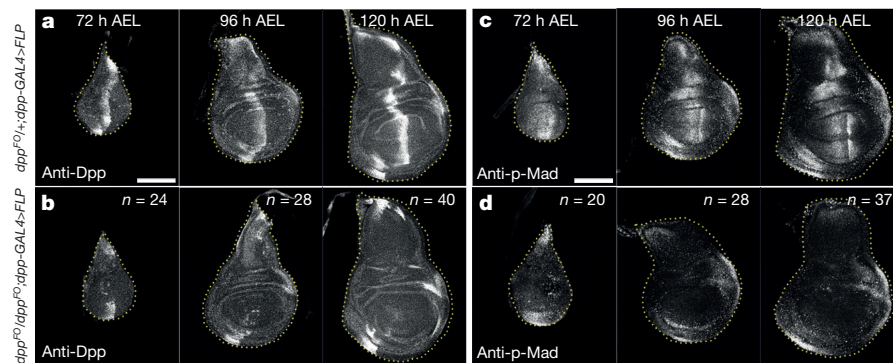
To test specifically the requirements for the Dpp morphogen gradient in disc growth, we used a disc-specific  $dpp-GAL4$  to drive expression of FLP in the compartmental stripe domain of  $dpp^{FO}/dpp^{FO}$  wing discs (Fig. 2b). Under these conditions Dpp protein was eliminated from the A/P compartmental boundary throughout the wing blade primordium ( $n = 40$ ; Fig. 2k, l and Extended Data Fig. 4e–k). Strikingly, however, the affected discs were grossly normal in both size and overall morphology (Fig. 2k–m). Some residual Dpp expression was detected in the posterior hinge, part of the ventral–anterior hinge and in some peripodial cells in which  $dpp-GAL4$  is not expressed (Fig. 2l, arrowheads, and Extended Data Fig. 4i–k). In addition, small clusters of cells expressing Dpp were frequently observed in proximity to the dorso–ventral boundary, perhaps due to reduced Gal4 expression ( $n = 27/40$ ; Fig. 2l, o and Extended Data Fig. 4i–k). However, consistent with the results of  $dpp^{d12}$  mosaic analyses (Fig. 1d–k and Extended Data Fig. 2), disruption

of the Dpp stripe caused only a mild growth defect relative to controls (7% reduction of area; Fig. 2k–m), without any obvious effect on cell proliferation (Fig. 2n–p). While we cannot rule out a contribution of residual Dpp to the growth of  $dpp^{FO}/dpp^{FO}$ ;  $dpp-GAL4/UAS-FLP$  wing discs, these experiments suggest that there is no instantaneous growth requirement for the canonical Dpp gradient centred on the A/P compartmental boundary of the wing blade primordium.

To address the kinetics of  $dpp$  disruption in  $dpp^{FO}/dpp^{FO}$ ;  $dpp-GAL4/UAS-FLP$  discs, we monitored Dpp expression and p-Mad activity at 72, 96 and 120 h AEL (Fig. 3). Compared with controls (Fig. 3a, c), the compartmental stripe of Dpp and its associated p-Mad activity gradient were disrupted in  $dpp^{FO}/dpp^{FO}$ ;  $dpp-GAL4/UAS-FLP$  wing discs dissected and fixed at 72 h AEL (Fig. 3b, d). The loss of Dpp and its activity gradient became more pronounced by 96 h AEL, and persisted as discs grew to a relatively normal size by 120 h AEL. Combined, these observations are consistent with the results of  $dpp^{d12}$  clonal analysis (Fig. 1d–k and Extended Data Fig. 2) and further support the conclusion that a continuous Dpp gradient centred on the compartmental stripe is not essential for proliferative growth in the third larval instar.

During wing disc development, p-Mad levels peak at the compartmental boundary and gradually diminish laterally (Fig. 4a)<sup>1,2</sup>. Upon phosphorylation, Mad translocates to the nucleus to regulate the transcriptional targets Sal, Omb and Brk (Fig. 4b–d), which in turn define the positions of the vein primordia visualized by anti-Delta (DI) staining (Fig. 4e)<sup>23</sup>. In agreement with this general model for wing patterning downstream of Dpp, removal of the compartmental Dpp stripe led to a pronounced loss of the p-Mad gradient ( $n = 37$ ; Fig. 4f) and





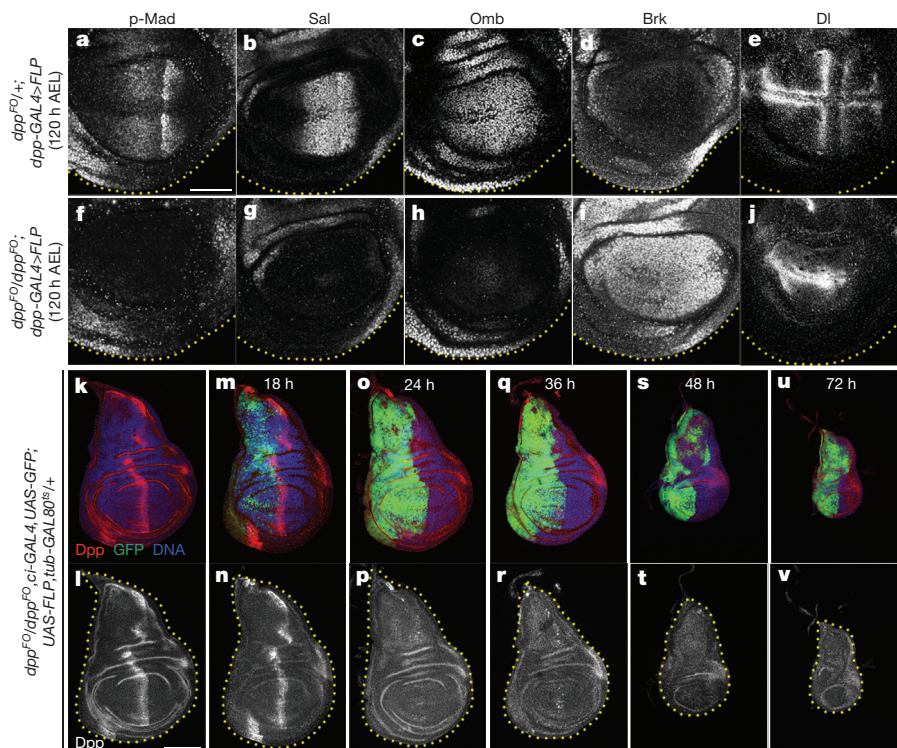
**Figure 3 | The Dpp gradient is not essential for growth in third instar wing discs.** **a, b,** Stripe Dpp expression is present in  $dpp^{F0/+}; dpp-GAL4/UAS-FLP$  controls (**a**) but eliminated from  $dpp^{F0}/dpp^{F0}; dpp-GAL4/UAS-FLP$  (**b**) wing discs during the third larval instar (72–120 h AEL).

the corresponding Sal ( $n = 25$ ; Fig. 4g) and Omb expression domains ( $n = 53$ ; Fig. 4h). In addition, Brk expression was de-repressed throughout the wing pouch ( $n = 25$ ; Fig. 4i) and patterning of presumptive vein territories was disrupted in discs devoid of the Dpp stripe ( $n = 20$ ; Fig. 4j). Intriguingly, in the course of these experiments we also noted that *dpp* disruption in dorsal cells led to a compartment-specific loss of p-Mad that was not rescued by Dpp of ventral origin (Extended Data Fig. 5). Likewise, *dpp*<sup>d12</sup> mutant clones adjacent to the dorsal–ventral border resulted in a compartment-specific loss of p-Mad staining (Fig. 1j, k and Extended Data Fig. 2g, h), suggesting either that Dpp protein is not able to cross the dorso–ventral boundary or that Dpp movement is directionally regulated along the disc A/P axis.

Taken together, our results confirm that *dpp* is a crucial regulator of wing disc pattern formation (Fig. 4a–j) but also demonstrate that the canonical morphogen gradient is not continuously required for growth and cell proliferation as the disc doubles in size during the third larval instar (Figs 2k–p, 3 and Extended Data Fig. 4f–k). This implies that the requirement for *dpp* in disc growth could either be fulfilled by earlier functions of the pathway<sup>24</sup> or by a cellular source of Dpp outside the classical stripe domain. An important possibility in  $dpp^{F0}/dpp^{F0}$ ;

**c, d,** The p-Mad activity gradient observed in  $dpp^{F0/+}; dpp-GAL4/UAS-FLP$  controls (**c**) is abolished in  $dpp^{F0}/dpp^{F0}; dpp-GAL4/UAS-FLP$  wing discs during the third larval instar (**d**). Scale bars, 100  $\mu$ m.

*dpp-GAL4/UAS-FLP* discs is that an initial burst of Dpp expression at the compartment boundary could precede FLP-mediated excision of the *dpp* locus and provide a sufficient stimulus to initiate early disc growth. Nevertheless, we found that global inactivation of *dpp* generated growth defects even at relatively late time points, consistent with a continuous requirement (Fig. 2h–j). To probe potential sources of Dpp expression outside of the stripe domain, we eliminated Dpp from defined spatial territories using the Gal4/UAS system. While loss of Dpp expression from whole discs or anterior cells alone elicited severe growth phenotypes, elimination of Dpp from posterior cells showed little or no effect (Extended Data Fig. 6a–d)<sup>25</sup>. Phenotypes caused by Dpp elimination with both *ap*- and *nub*-GAL4 were consistent with the interpretation that Dpp produced by anterior cells is necessary for wing disc growth (Extended Data Fig. 6e, f), even though the compartmental stripe itself was not essential (Fig. 2l, o). To assess the temporal requirements for *dpp* in anterior cells, we used *GAL80<sup>ts</sup>* to repress *ci-GAL4* activity (Fig. 4k–v and Extended Data Fig. 7)<sup>26</sup>. In wing discs from  $dpp^{F0}/dpp^{F0}; ci-GAL4, UAS-GFP; UAS-FLP, tub-GAL80^{ts}/+$  larvae, Dpp expression was generally disrupted within 24 h of removing *GAL80*-mediated repression of Gal4 ( $n = 34/40$ ; Fig. 4o, p). Inactivation



**Figure 4 | The Dpp stripe is crucial for wing pattern formation.** **a–j,** The p-Mad activity gradient and BMP-dependent gene expression domains observed in controls (**a–e**) are severely disrupted following loss of the Dpp gradient in  $dpp^{F0}/dpp^{F0}; dpp-GAL4/UAS-FLP$  wing discs (**f–j**). Scale bar, 50  $\mu$ m. **k–v,** Dpp elimination from the anterior compartment of wing discs at a series of time points before dissection. In all cases,  $dpp^{F0}/dpp^{F0}; ci-GAL4, UAS-GFP; UAS-FLP, tub-GAL80^{ts}/+$  wing discs were stained for anti-Dpp (red) and DNA (blue). Green fluorescent protein (GFP; green) indicates de-repressed Gal4 activity. Compared with controls (**k, l**), normal Dpp expression is maintained for 18 h after temperature shift (**m, n**;  $n = 33$ ) but mostly eliminated within 24 h (**o, p**;  $n = 34/40$ ). Normal Dpp staining is lost after temperature shifts at 36 (**q, r**;  $n = 33$ ), 48 (**s, t**;  $n = 80$ ) and 72 (**u, v**;  $n = 43$ ) h before dissection. Scale bar, 100  $\mu$ m.

of anterior *dpp* at earlier larval time points through temperature shifts caused the most severe growth defects (Fig. 4s–v), but even late inactivation resulted in mildly reduced disc size (Fig. 4q, r). Although these experiments offer limited temporal resolution, the results indicate that Dpp produced within the *ci-GAL4* expression domain is required for wing disc growth throughout development. By inference, this would indicate that Dpp expressed within the anterior compartment is sufficient to sustain homogenous disc growth in the absence of the canonical morphogen gradient during the third larval instar.

*Drosophila* Dpp has long served as a central paradigm for understanding how morphogens regulate growth and patterning during animal development. To date, several distinct models have been proposed for imaginal disc growth control by the Dpp activity gradient<sup>7–13</sup>. We directly tested the requirements for a gradient of Dpp signalling during wing disc development. Consistent with established *dpp* mutant phenotypes and the critical temporal requirement for Dpp during early development<sup>24</sup>, eliminating Dpp throughout the disc at early larval stages caused severe growth defects (Fig. 2h). Surprisingly, however, abrogation of the compartment-boundary-centred Dpp signalling gradient did not disrupt active cell proliferation and elicited only mild growth defects during the third larval instar (Figs 2k–p, 3 and Extended Data Fig. 4f–k). In summary, while Dpp is clearly required for disc growth, we propose that the classical Dpp morphogen gradient primarily regulates pattern formation and is not continuously required to drive proliferative growth in the latter half of larval development. These findings suggest dynamic spatial and temporal requirements for Dpp. Expanding on our results, we speculate that other morphogen systems may utilize a similar strategy to coordinate growth and patterning during organ and appendage development.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 13 March; accepted 14 September 2015.**

**Published online 9 November 2015.**

1. Raftery, L. A. & Umulis, D. M. Regulation of BMP activity and range in *Drosophila* wing development. *Curr. Opin. Cell Biol.* **24**, 158–165 (2012).
2. Affolter, M. & Basler, K. The Decapentaplegic morphogen gradient: from pattern formation to growth regulation. *Nature Rev. Genet.* **8**, 663–674 (2007).
3. Lecuit, T. *et al.* Two distinct mechanisms for long-range patterning by Decapentaplegic in the *Drosophila* wing. *Nature* **381**, 387–393 (1996).
4. Nellen, D., Burke, R., Struhl, G. & Basler, K. Direct and long-range action of a DPP morphogen gradient. *Cell* **85**, 357–368 (1996).
5. Akiyama, T. & Gibson, M. C. Morphogen transport: theoretical and experimental controversies. *Wiley Interdiscip. Rev. Dev. Biol.* **4**, 99–112 (2015).
6. Restrepo, S., Zartman, J. J. & Basler, K. Coordination of patterning and growth by the morphogen DPP. *Curr. Biol.* **24**, R245–R255 (2014).
7. Rogulja, D. & Irvine, K. D. Regulation of cell proliferation by a morphogen gradient. *Cell* **123**, 449–461 (2005).

8. Rogulja, D., Rauskolb, C. & Irvine, K. D. Morphogen control of wing growth through the Fat signaling pathway. *Dev. Cell* **15**, 309–321 (2008).
9. Schwank, G., Restrepo, S. & Basler, K. Growth regulation by Dpp: an essential role for Brinker and a non-essential role for graded signaling levels. *Development* **135**, 4003–4013 (2008).
10. Schwank, G. *et al.* Antagonistic growth regulation by Dpp and Fat drives uniform cell proliferation. *Dev. Cell* **20**, 123–130 (2011).
11. Schwank, G., Yang, S. F., Restrepo, S. & Basler, K. Comment on “Dynamics of Dpp signaling and proliferation control”. *Science* **335**, 401 (2012).
12. Wartlick, O. *et al.* Dynamics of Dpp signaling and proliferation control. *Science* **331**, 1154–1159 (2011).
13. Day, S. J. & Lawrence, P. A. Measuring dimensions: the regulation of size and shape. *Development* **127**, 2977–2987 (2000).
14. Lecuit, T. & Le Goff, L. Orchestrating size and shape during morphogenesis. *Nature* **450**, 189–192 (2007).
15. Schwank, G. & Basler, K. Regulation of organ growth by morphogen gradients. *Cold Spring Harb. Perspect. Biol.* **2**, a001669 (2010).
16. Kicheva, A., Bollenbach, T., Wartlick, O., Jülicher, F. & Gonzalez-Gaitan, M. Investigating the principles of morphogen gradient formation: from tissues to cells. *Curr. Opin. Genet. Dev.* **22**, 527–532 (2012).
17. Müller, P., Rogers, K. W., Yu, S. R., Brand, M. & Schier, A. F. Morphogen transport. *Development* **140**, 1621–1638 (2013).
18. Spencer, F. A., Hoffmann, F. M. & Gelbart, W. M. Decapentaplegic: a gene complex affecting morphogenesis in *Drosophila melanogaster*. *Cell* **28**, 451–461 (1982).
19. Zecca, M., Basler, K. & Struhl, G. Sequential organizing activities of engrailed, hedgehog and decapentaplegic in the *Drosophila* wing. *Development* **121**, 2265–2278 (1995).
20. St Johnston, R. D. *et al.* Molecular organization of the decapentaplegic gene in *Drosophila melanogaster*. *Genes Dev.* **4**, 1114–1127 (1990).
21. Morata, G. & Ripoll, P. Minutes: mutants of *Drosophila* autonomously affecting cell division rate. *Dev. Biol.* **42**, 211–221 (1975).
22. Domínguez, M., Wasserman, J. D. & Freeman, M. Multiple functions of the EGF receptor in *Drosophila* eye development. *Curr. Biol.* **8**, 1039–1048 (1998).
23. Blair, S. S. Wing vein patterning in *Drosophila* and the analysis of intercellular signaling. *Annu. Rev. Cell Dev. Biol.* **23**, 293–319 (2007).
24. Paul, L. *et al.* Dpp-induced Egfr signaling triggers postembryonic wing development in *Drosophila*. *Proc. Natl Acad. Sci. USA* **110**, 5058–5063 (2013).
25. Foronda, D., Pérez-Garijo, A. & Martín, F. A. Dpp of posterior origin patterns the proximal region of the wing. *Mech. Dev.* **126**, 99–106 (2009).
26. McGuire, S. E., Le, P. T., Osborn, A. J., Matsumoto, K. & Davis, R. L. Spatiotemporal rescue of memory dysfunction in *Drosophila*. *Science* **302**, 1765–1768 (2003).

**Acknowledgements** We thank G. Struhl and K. Wharton for extensive discussion and suggestions, S. Kondo for technical advice with CRISPR, K. Irvine, W. Deng and the Bloomington Stock Center for fly stocks, and R. Barrio, G. Pflugfelder, A. Teleman and the Developmental Studies Hybridoma Bank for antibodies. We thank K. Marr and L. Ellington for a critical reading of the manuscript, L. Gutchevsky for administrative support, and members of the Gibson laboratory for discussions and advice. This work was supported by funding from the Stowers Institute for Medical Research.

**Author Contributions** T.A. and M.C.G. conceived the project, designed the experiments and wrote the manuscript. T.A. performed the experiments and analysed the data.

**Additional Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.C.G. (MG2@stowers.org).



## METHODS

**Dpp antibody production.** *dpp* complementary DNA was cloned into *pET-DEST42* (Invitrogen) using the GATEWAY system. The expression of Dpp-His6 was induced by the addition of 1 mM isopropylthiogalactoside (IPTG) at an OD<sub>600 nm</sub> of 0.5 in 1 l of LB media at 37 °C. After a 3 h induction, cells were harvested, resuspended in 50 ml of denaturing buffer (50 mM Tris-HCl, 1 M NaCl, 20 mM imidazole, 6 M urea, and 0.1% Triton X-100, pH 7.5) and disrupted by sonication. The cell lysate was applied to a Ni Sepharose column for protein purification (GE Health Care Life Sciences). The column was washed (50 mM Tris-HCl, 1 M NaCl, 6 M urea, and 0.1% Triton X-100, pH 7.5) and DPP-His6 proteins were eluted with 2 ml of elution buffer (50 mM Tris-HCl, 1 M NaCl, 500 mM imidazole, 6 M urea, and 0.1% Triton X-100, pH 7.5). The purified protein was renatured by step-wise reduction of urea concentration. The soluble fraction of the purified Dpp-His6 protein was used to immunize two rabbits pre-screened for low serum immunoreactivity against imaginal discs, one of which produced suitable antibodies. Rabbit anti-sera against Dpp-His6 were affinity purified using the Dpp prodomain (amino acids 1–456).

**Clonal analyses of *dpp*<sup>d12</sup> mutant cells in wing discs.** All flies were maintained using standard cornmeal media at 25 °C. *dpp*<sup>d12</sup> is a strong hypomorphic allele that carries a chromosomal inversion in the *dpp*<sup>disc</sup> enhancer region<sup>20</sup>. Mutant clones were generated by the FLP-FRT method<sup>27</sup> and marked by the absence of GFP according to the following schemes. Control cross: *y,w,hs-FLP; ubi-GFP,FRT40A/CyO* × *w; FRT40A*; experimental cross: *y,w,hs-FLP; ubi-GFP,FRT40A/CyO* × *w; dpp*<sup>d12</sup>,*FRT40A/CyO,twi-GAL4,UAS-GFP*.

*dpp*<sup>d12</sup> clones in a *Minute* background<sup>21,22</sup> were generated as follows. Control cross: *y,w,hs-FLP; ubi-GFP,M(2)25A,FRT40A/CyO* × *w; FRT40A*; experimental cross: *y,w,hs-FLP; ubi-GFP,M(2)25A,FRT40A/CyO* × *w; dpp*<sup>d12</sup>,*FRT40A/CyO,twi-GAL4,UAS-GFP*. Clones were induced at 37 °C for 2 h at 48 h AEL (standard) or 72 h AEL (*Minute*).

**Western blot analysis.** Wing discs were homogenized in SDS sample buffer and boiled. Protein samples were subjected to 4–20% Mini-PROTEAN TGX gel (Bio-Rad) or 10% SDS-PAGE gel electrophoresis and then analysed by western blot with SuperSignal (Thermo). Rabbit anti-Dpp (1:1,000), mouse anti- $\alpha$ -tubulin (1:1,000, Sigma) and horseradish peroxidase (HRP)-conjugated secondary antibodies (1:10,000, Jackson ImmunoResearch) were used for detection.

**Generation of an FLP/FRT-mediated conditional *dpp*-null allele.** *pBfV-U6.2* (ref. 28) was used for making sgRNA DNA constructs. The following primers were annealed and cloned into the BbsI site of *pBfV-U6.2*. sgRNA construct 1: 5'-CTT CGGTTCCGATGTGGACCGAA-3', 5'-AACTTCCGGTCCACATCCGAA CC-3'; sgRNA construct 2: 5'-CTTCGGACAGAAGGATCTAGGGAT-3', 5'-AA ACATCCCTAGATCCTTCTGTCC-3'.

To generate a *ubi-GFP* selection cassette, the 1,760 bp promoter region of *ubiquitin-63E* was amplified from *w*<sup>1118</sup> genomic DNA using 5'-GGCGGCGAA TTCATCAGTACTGTCCAAAATCGAAATCGCCGAACCG-3' and 5'-GGC GGCGGTACCTTTGGATTATTCTGCGGGAAGAAAATAGAGATGTGG-3' primers with EcoRI and KpnI sites at the 5' and 3' ends, respectively (restriction enzyme sites in primers are in bold). Then, the PCR product was cloned into the EcoRI and KpnI sites of *pH-Stinger*<sup>29</sup>.

The Gibson assembly technique (NEB) was employed to obtain a donor DNA for CRISPR-Cas9-mediated homologous recombination. First, five PCR fragments were prepared using the following primer sets. Fragment 1 (*pHS298* vector): 5'-CCGGGTACCGAGCTCGAA-3', 5'-GGATCCTCTAGAGTTCGACCTG-3'; fragment 2 (left arm homology-FRT 5'): 5'-AGGTCGACTCTAGAGGATCCCG AAAGATCCCTTTGCGC-3', 5'-TTATGATATCGAAGTTCTTACTATTCTA GAGAATAGGAACCTCGGAATAGGAACCTCGAATGGAATCGCGTTCGT ATTCCATCAATCC-3'; fragment 3 (*loxP* 5'-*ubi-GFP* selection cassette-*loxP* 3'): 5'-TAGGAACCTCGATATCAATCTCGTATAGCATACATTATACGAA GTTATTGCGCAAGCTTGGGCTGCATCACGTAATAAGTGTGCGTTG-3', 5'-ATGTGGACCGATAACTTCGTATAATGTATGCTATACGAAGTTATTTA ACTTACATACATAGTAATGATCGGCTAAATGGTATGGC-3'; fragment 4 (*dpp-FRT* 3'): 5'-ACGAAGTTATCGGTCCACATCCGAACCC-3', 5'-AGGAT CTAGGGAAGTTCCATATCTTCTAGAGAATAGGAACCTCGGAATAGGAA CTTCCATATGGATCGGCAGGTATGCAATCGCTTAG-3'; fragment 5 (right arm homology): 5'-TAGGAACCTCCCTAGATCCTTCTGTCTCG-3', 5'-ATT CGAGCTCGGTACCCGGGCGGGAATGCTCTTACAGTC-3'.

After the PCR reaction, all fragments were purified using ZymoClean Gel DNA Recovery kit (Zymo research) and subjected to Gibson assembly (NEB).

After confirming the DNA sequences, these plasmid DNAs were mixed, precipitated by ethanol precipitation and dissolved in nuclease-free water with food dye at 250 ng  $\mu$ l<sup>-1</sup> for each DNA construct. The DNA mixture was injected into the posterior side of embryos obtained from a cross of *w*<sup>1118</sup> and *y,cho,v;*

*attP40{nos-Cas9}/CyO* (ref. 28). Transgenic flies were first selected by GFP expression and were further confirmed by Southern blots, PCR and DNA sequencing. **Genomic Southern blot analysis.** Genomic DNAs from *w*<sup>1118</sup> and *w; dpp*<sup>FO-GFP</sup>/CyO adults were prepared as described previously<sup>30</sup>, digested with ClaI at 37 °C for 4 h, and subjected to 0.7% agarose gel electrophoresis. After electrophoresis, Southern blotting was performed using a standard protocol described previously<sup>31</sup>. A DIG-labelled GFP probe was generated using DIG DNA labelling kit (Roche). After hybridization at 65 °C overnight, hybridized DNA fragments were visualized via alkaline phosphatase conjugated anti-Dig (1:5,000, Roche).

**PCR analysis of FLP/FRT-mediated *dpp* knockdown.** Fifty wing discs from *w; dpp*<sup>FO</sup>/*dpp*<sup>FO</sup> and *w; dpp*<sup>FO</sup>/*dpp*<sup>FO</sup>; *dpp-GAL4/UAS-FLP* were collected to obtain genomic DNAs using Maxwell 16 system (Promega). Then, PCR was carried out using 5'-CCACCGATCCGCTTATCGGAGG-3' and 5'-CGCCGCTTCAGCTTCTCGTCG-3' primers.

**Heat shock.** Control cross: *y,w,hs-FLP; dpp*<sup>FO</sup>/CyO,*sChFP* × *w*<sup>1118</sup>; experimental cross: *y,w,hs-FLP; dpp*<sup>FO</sup>/CyO,*sChFP* × *y,w,hs-FLP; dpp*<sup>FO</sup>/CyO,*sChFP*. *dpp* mutant cells were induced by heat shock at 37 °C for 30 min at 48, 72 or 96 h AEL (ref. 32). **GAL4/UAS. *dpp-GAL4* (*dpp*<sup>blk</sup>-*GAL4*;** *GAL4* expression is controlled by a partial *dpp*<sup>disc</sup> enhancer<sup>33</sup>, *en-GAL4* (ref. 34), *ci-GAL4* (ref. 35), *ap-GAL4* (ref. 36) and *nub-GAL4* (ref. 36) were used to induce FLP to eliminate Dpp expression from specific regions of wing discs using the Gal4/UAS system (ref. 37). The efficiency of FLP/FRT-mediated recombination was monitored by using *Act5c(-FRT)lacZ* (ref. 38).

***dpp-GAL4.*** Control cross: *w; dpp*<sup>FO</sup>/CyO,*sChFP*; *UAS-FLP/TM6C* × *w; dpp-GAL4/TM6B*; experimental cross: *w; dpp*<sup>FO</sup>/CyO,*sChFP*; *UAS-FLP/TM6C* × *w; dpp*<sup>FO</sup>/CyO,*sChFP*; *dpp-GAL4/TM6C*.

***dpp-GAL4, Act5c(-FRT)lacZ.*** Control cross: *w; dpp*<sup>FO</sup>/CyO,*sChFP*; *UAS-FLP,Act5c(-FRT)lacZ/TM6C* × *dpp-GAL4/TM6B*; experimental cross: *w; dpp*<sup>FO</sup>/CyO,*sChFP*; *UAS-FLP,Act5c(-FRT)lacZ/TM6C* × *w; dpp*<sup>FO</sup>/CyO,*sChFP*; *dpp-GAL4/TM6C*.

***en-GAL4,ci-GAL4.*** Control cross: *w; dpp*<sup>FO</sup>/CyO,*sChFP*; *UAS-FLP/TM6C* × *w; en-GAL4,ci-GAL4* (on II); experimental cross: *w; dpp*<sup>FO</sup>/CyO,*sChFP*; *UAS-FLP/TM6C* × *w; dpp*<sup>FO</sup>,*en-GAL4,ci-GAL4/CyO,twi-GAL4,UAS-GFP*.

***en-GAL4.*** Control cross: *w; dpp*<sup>FO</sup>/CyO,*sChFP*; *UAS-FLP/TM6C* × *w; en-GAL4* (on II); experimental cross: *w; dpp*<sup>FO</sup>/CyO,*sChFP*; *UAS-FLP/TM6C* × *w; dpp*<sup>FO</sup>,*en-GAL4/CyO,twi-GAL4,UAS-GFP*.

***ci-GAL4.*** Control cross: *w; dpp*<sup>FO</sup>/CyO,*sChFP*; *UAS-FLP/TM6C* × *w; ci-GAL4* (on II); experimental cross: *w; dpp*<sup>FO</sup>/CyO,*sChFP*; *UAS-FLP/TM6C* × *w; dpp*<sup>FO</sup>,*ci-GAL4/CyO,twi-GAL4,UAS-GFP*.

***ci-GAL4, tub-GAL80<sup>ts</sup>.*** Control cross: *w; dpp*<sup>FO</sup>/CyO,*sChFP*; *UAS-FLP,tub-GAL80<sup>ts</sup>/TM6C* × *ci-GAL4,UAS-GFP* (on II); experimental cross: *w; dpp*<sup>FO</sup>/CyO,*sChFP*; *UAS-FLP,tub-GAL80<sup>ts</sup>/TM6C* × *w; dpp*<sup>FO</sup>,*ci-GAL4,UAS-GFP* (on II).

***ap-GAL4.*** Control cross: *w; dpp*<sup>FO</sup>/CyO,*sChFP*; *UAS-FLP/TM6C* × *w; ap-GAL4/T2;3*; experimental cross: *w; dpp*<sup>FO</sup>/CyO,*sChFP*; *UAS-FLP/TM6C* × *w; dpp*<sup>FO</sup>,*ap-GAL4/CyO,twi-GAL4,UAS-GFP*.

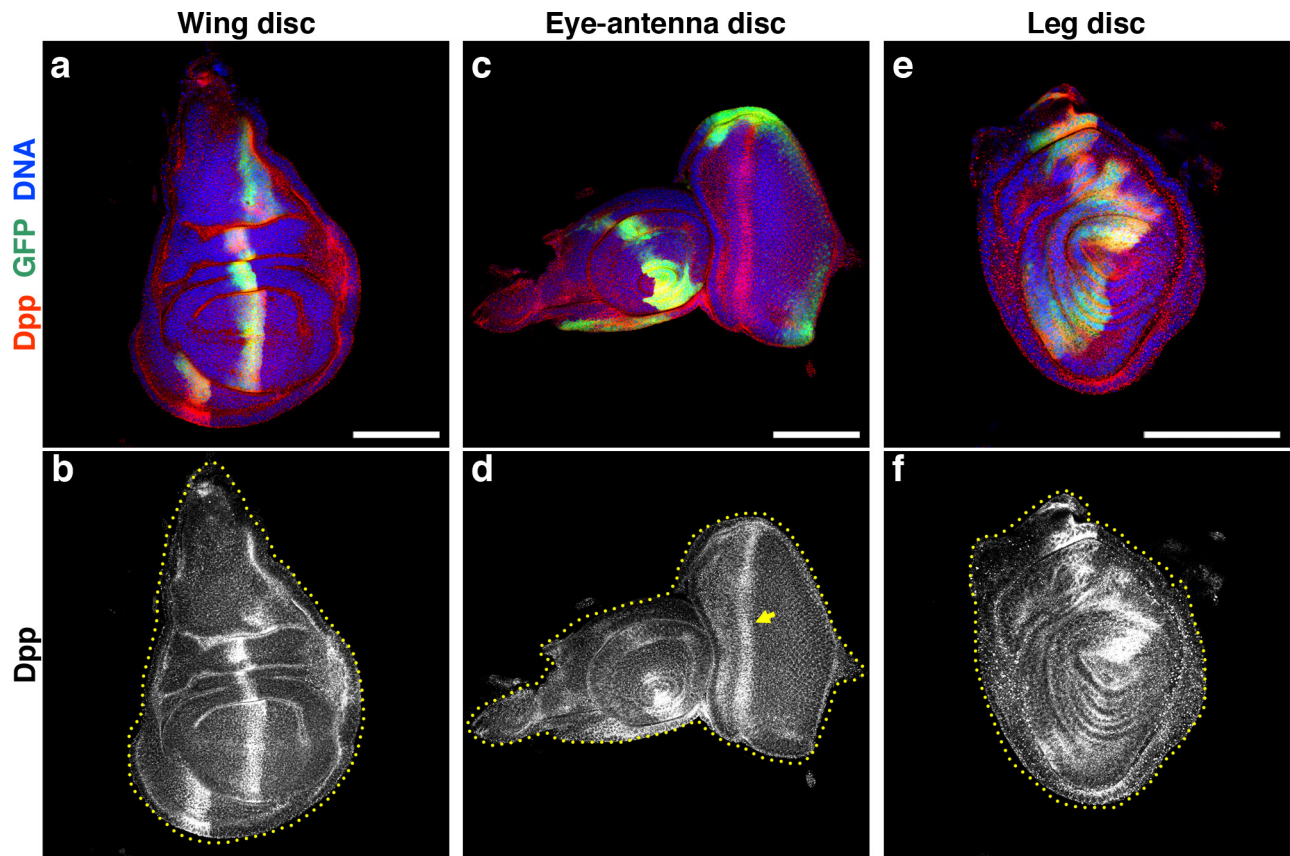
***nub-GAL4.*** Control cross: *w; dpp*<sup>FO</sup>/CyO,*sChFP*; *UAS-FLP/TM6C* × *w; nub-GAL4* (on II); experimental cross: *w; dpp*<sup>FO</sup>/CyO,*sChFP*; *UAS-FLP/TM6C* × *w; dpp*<sup>FO</sup>,*nub-GAL4/CyO,twi-GAL4,UAS-GFP*.

**Immunohistochemistry and imaging.** Immunostaining of wing discs was carried out as previously described<sup>39</sup> with some modifications. Rabbit anti-Dpp (1:100), mouse anti-p-H3 (1:1,000, Millipore), mouse anti-Dlg (1:500, DSHB), mouse anti- $\beta$ -galactosidase (Z3781, 1:200, Promega), rabbit anti-pSmad3 (EP823Y, 1:1,000, Epitomics)<sup>40,41</sup>, rat anti-Sal (1:200)<sup>42</sup>, rabbit anti-Omb (1:1,000)<sup>43</sup>, guinea pig anti-Brk (1:500)<sup>44</sup>, mouse anti-Dl (C594.9B, 1:500, DSHB)<sup>45</sup>, and Alexa-conjugated secondary antibodies (1:500, Invitrogen) were used for this study. All primary antibodies were diluted in Can Get Signal Immunostain Solution B (TOYOBO). DNA was visualized using Hoechst 33342 (1:1,000; Thermo Scientific). Images of wing discs were obtained using a Leica TCS SP5 confocal microscope.

**Image analysis.** Sizes of wild-type wing discs at 72 h (*n* = 61), 96 h (*n* = 57) and 120 h AEL (*n* = 84) were measured using the 'measure' function of FIJI and analysed using Student's *t*-test. The same method was used to measure discs of the following genotypes: *dpp*<sup>d12</sup>/*dpp*<sup>d12</sup> (*n* = 25), *hs-FLP; dpp*<sup>FO</sup>/*dpp*<sup>FO</sup> (heat shocked at 48 h (*n* = 23), 72 h (*n* = 22) and 96 h AEL (*n* = 26)), *dpp*<sup>FO</sup>/+, *dpp-GAL4/UAS-FLP* (*n* = 187) and *dpp*<sup>FO</sup>/*dpp*<sup>FO</sup>; *dpp-GAL4/UAS-FLP* (*n* = 252). Mitotic index was determined by dividing the numbers of mitotic cells (p-H3-positive cells) by total cell numbers (visualized by anti-Dlg staining) in a 30.03  $\mu$ m square in the anterior-ventral wing pouch region. The 'find maxima' function of FIJI was used to automatically count total cell numbers. No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

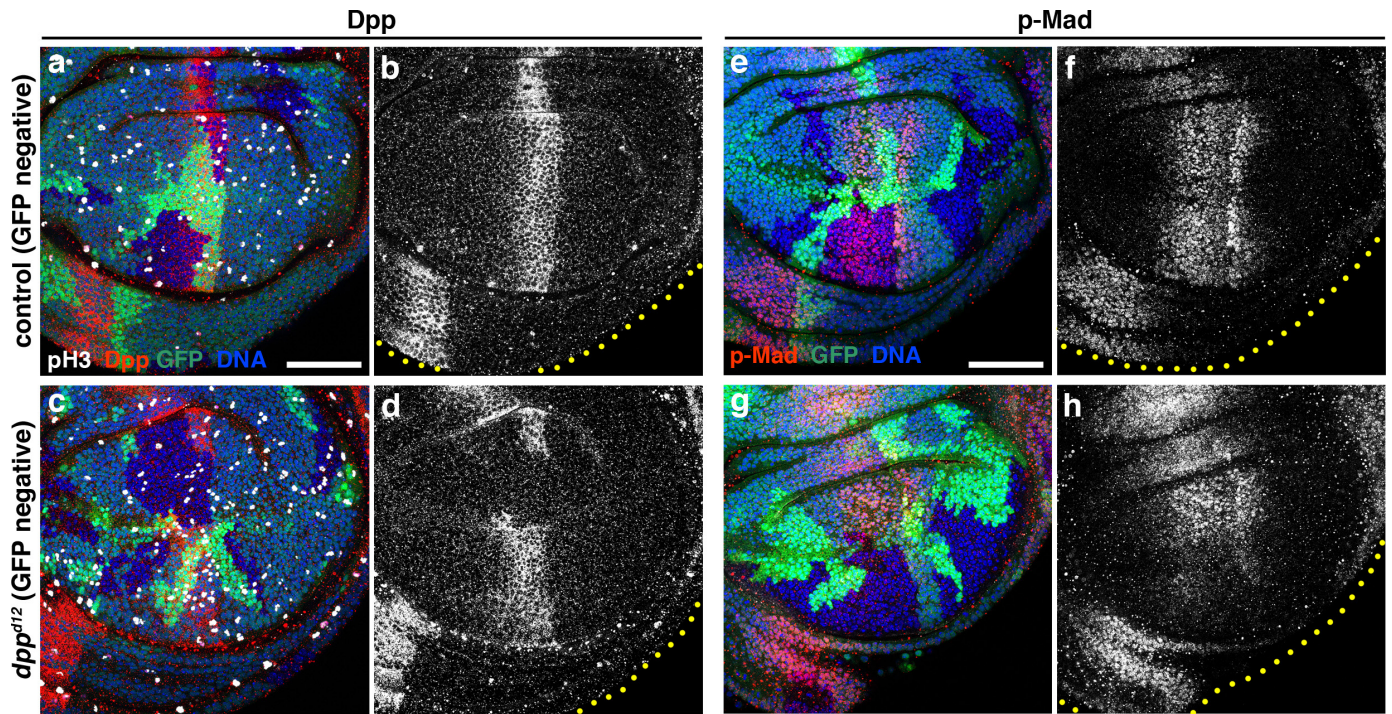


27. Xu, T. & Rubin, G. M. Analysis of genetic mosaics in developing and adult *Drosophila* tissues. *Development* **117**, 1223–1237 (1993).
28. Kondo, S. & Ueda, R. Highly improved gene targeting by germline-specific Cas9 expression in *Drosophila*. *Genetics* **195**, 715–721 (2013).
29. Barolo, S., Carver, L. A. & Posakony, J. W. GFP and  $\beta$ -galactosidase transformation vectors for promoter/enhancer analysis in *Drosophila*. *Biotechniques* **29**, 726, 728, 730, 732 (2000).
30. Huang, A. M., Rehm, E. J. & Rubin, G. M. Quick preparation of genomic DNA from *Drosophila*. *Cold Spring Harb. Protoc.* **2009**, pdb.prot5198 (2009).
31. Brown, T. Southern blotting. *Curr. Protoc. Mol. Biol.* **Chapter 2**, Unit 2.9A (2001).
32. Golic, K. G. & Lindquist, S. The FLP recombinase of yeast catalyzes site-specific recombination in the *Drosophila* genome. *Cell* **59**, 499–509 (1989).
33. Staehling-Hampton, K., Jackson, P. D., Clark, M. J., Brand, A. H. & Hoffmann, F. M. Specificity of bone morphogenetic protein-related factors: cell fate and gene expression changes in *Drosophila* embryos induced by decapentaplegic but not 60A. *Cell Growth Differ.* **5**, 585–593 (1994).
34. Johnson, R. L., Grenier, J. K. & Scott, M. P. patched overexpression alters wing disc size and pattern: transcriptional and post-transcriptional effects on hedgehog targets. *Development* **121**, 4161–4170 (1995).
35. Croker, J. A., Ziegenhorn, S. L. & Holmgren, R. A. Regulation of the *Drosophila* transcription factor, Cubitus interruptus, by two conserved domains. *Dev. Biol.* **291**, 368–381 (2006).
36. Calleja, M., Moreno, E., Pelaz, S. & Morata, G. Visualization of gene expression in living adult *Drosophila*. *Science* **274**, 252–255 (1996).
37. Brand, A. H. & Perrimon, N. Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. *Development* **118**, 401–415 (1993).
38. Struhl, G. & Basler, K. Organizing activity of wingless protein in *Drosophila*. *Cell* **72**, 527–540 (1993).
39. Akiyama, T. *et al.* Dally regulates Dpp morphogen gradient formation by stabilizing Dpp on the cell surface. *Dev. Biol.* **313**, 408–419 (2008).
40. Akiyama, T., Marqués, G. & Wharton, K. A. A large bioactive BMP ligand with distinct signaling properties is produced by alternative proconvertase processing. *Sci. Signal.* **5**, ra28 (2012).
41. Dejima, K., Kanai, M. I., Akiyama, T., Levings, D. C. & Nakato, H. Novel contact-dependent bone morphogenetic protein (BMP) signaling mediated by heparan sulfate proteoglycans. *J. Biol. Chem.* **286**, 17103–17111 (2011).
42. Barrio, R. & de Celis, J. F. Regulation of *spalt* expression in the *Drosophila* wing blade in response to the Decapentaplegic signaling pathway. *Proc. Natl Acad. Sci. USA* **101**, 6021–6026 (2004).
43. Shen, J., Dahmann, C. & Pflugfelder, G. O. Spatial discontinuity of optomotor-blind expression in the *Drosophila* wing imaginal disc disrupts epithelial architecture and promotes cell sorting. *BMC Dev. Biol.* **10**, 23 (2010).
44. Doumpas, N. *et al.* Brk regulates wing disc growth in part via repression of Myc expression. *EMBO Rep.* **14**, 261–268 (2013).
45. Bangi, E. & Wharton, K. Dpp and Gbb exhibit different effective ranges in the establishment of the BMP activity gradient critical for *Drosophila* wing patterning. *Dev. Biol.* **295**, 178–193 (2006).



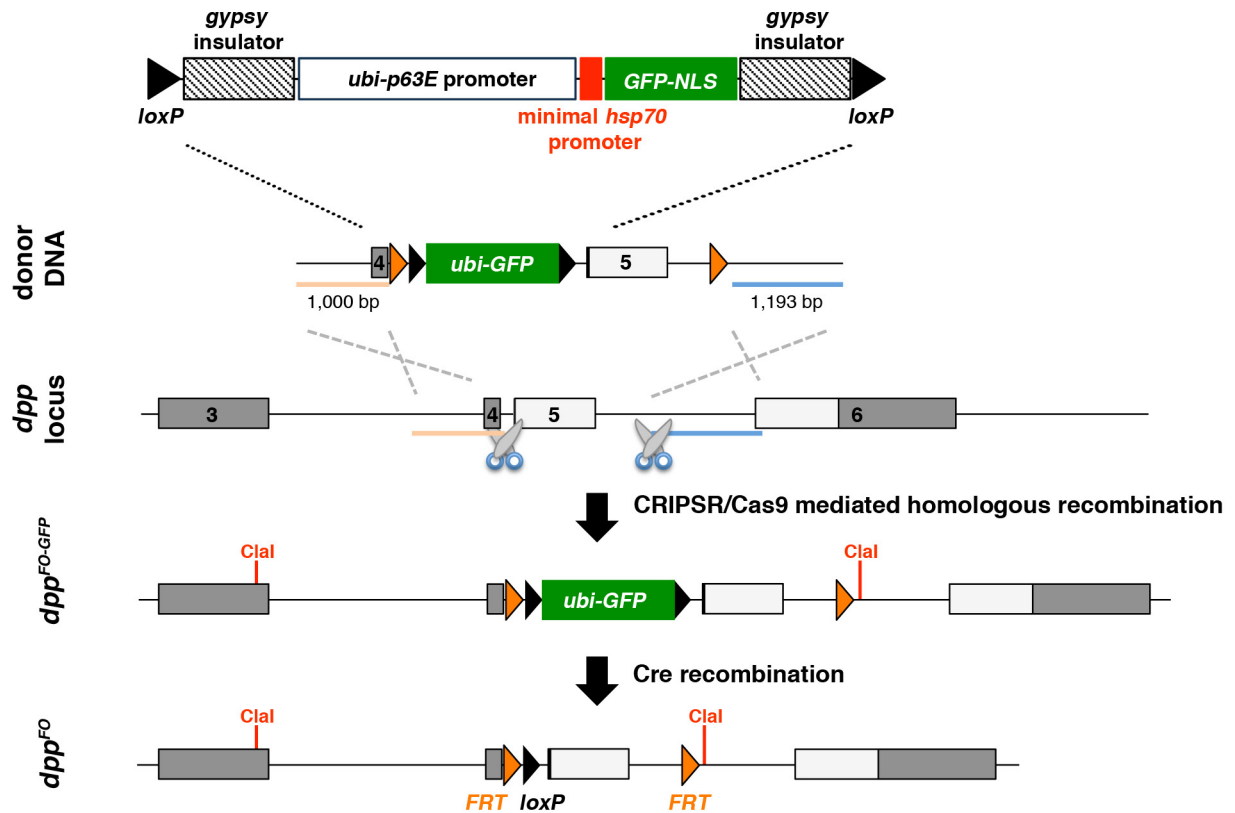
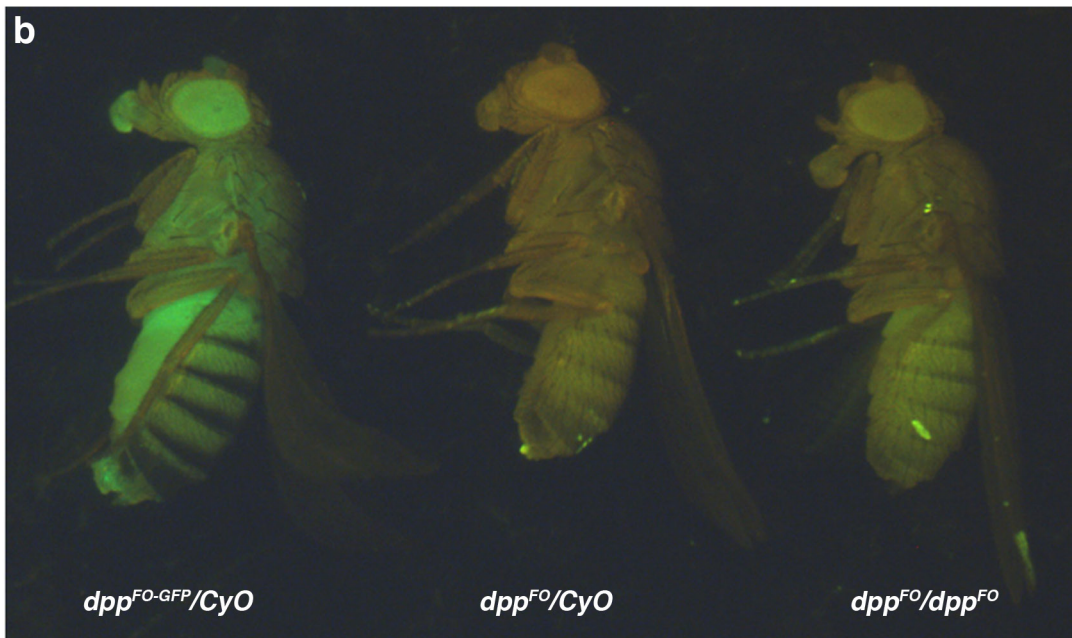
**Extended Data Figure 1 | Endogenous Dpp expression in imaginal discs.** a–f, Wing (a, b), eye–antenna (c, d), and leg (e, f) imaginal discs from *UAS-GFP/+; dpp-GAL4/+* larvae are dissected and stained with anti-Dpp antibody. GFP (green) indicates *dpp-GAL4*-expressing cells.

Note that *dpp-GAL4* is not expressed in the morphogenetic furrow of the third instar eye–antenna disc (arrow in d). Dotted lines show outlines of imaginal discs. Blue: DNA. Scale bars, 100  $\mu$ m. Anterior is left.



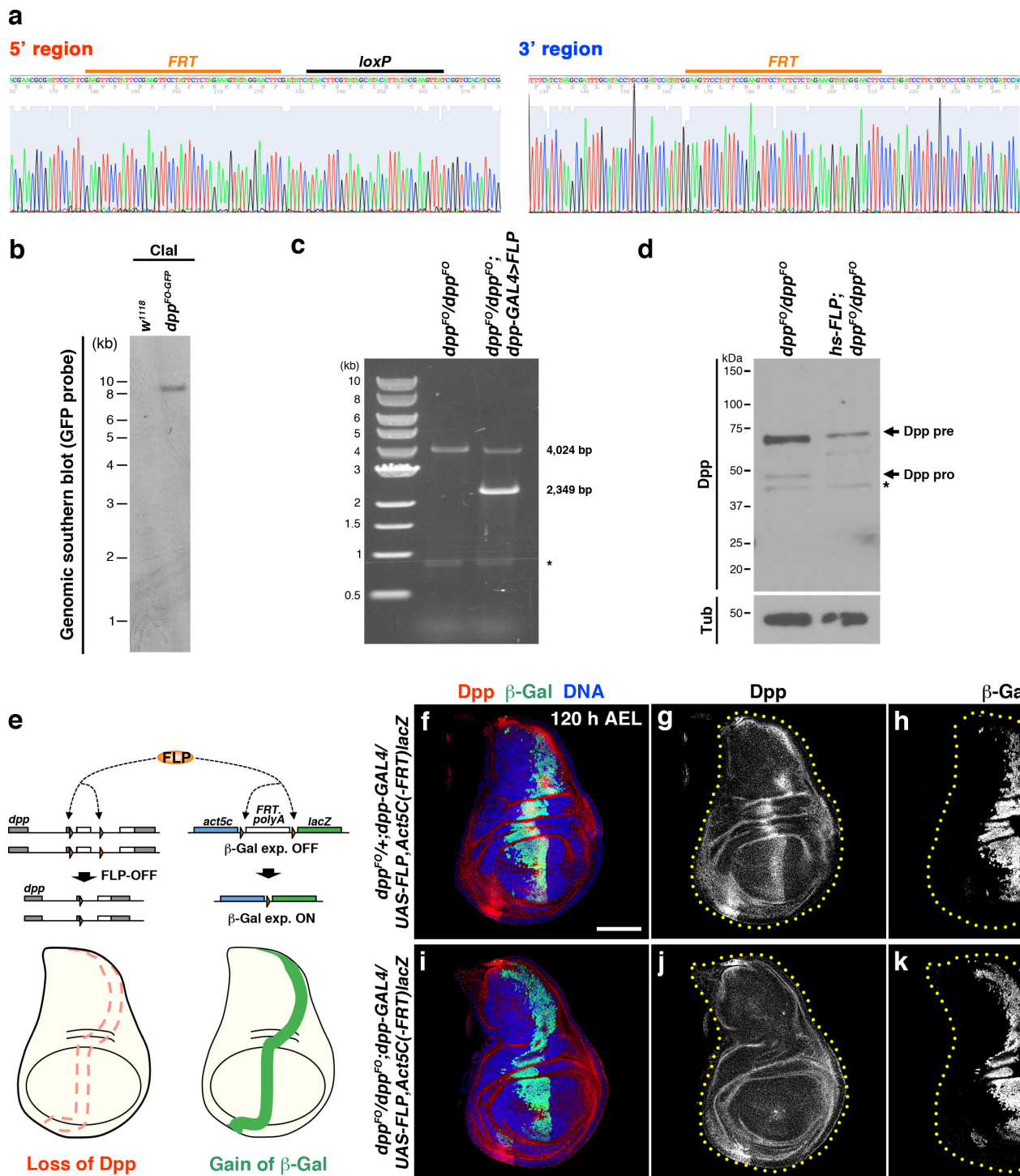
Extended Data Figure 2 | Dpp and p-Mad expression in *dpp<sup>d12</sup>* clones. a–h, Control (a, b, e, f) and *dpp<sup>d12</sup>* mutant (c, d, g, h) clones are stained with anti-Dpp (a–d) and anti-p-Mad (e–h) antibodies. Clones are marked by the absence of GFP (green). Disc boundaries are indicated by dotted lines. Scale bars, 50  $\mu$ m. Anterior is to the left.



**a****b**

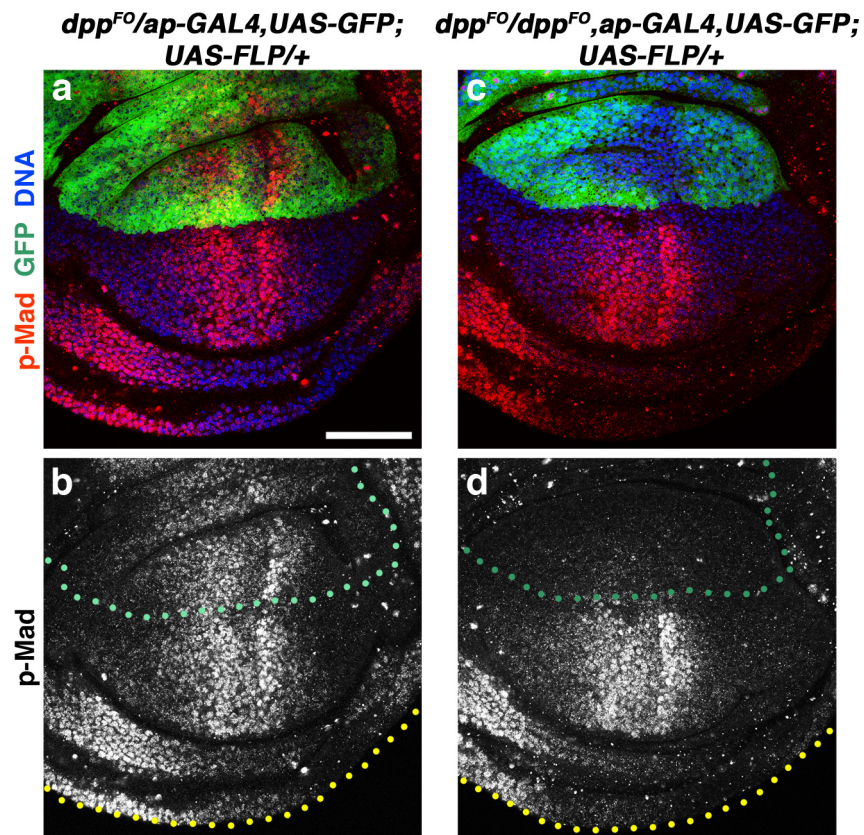
**Extended Data Figure 3 | FLP/FRT-mediated conditional *dpp*-null allele.** **a**, A flowchart describing the establishment of an FLP/FRT-mediated conditional *dpp*-null allele. Grey and white boxes indicate untranslated region (UTR) and *dpp* coding sequences, respectively. *FRT* sequences flank the first coding exon (exon 5). Since this exon contains the

*Dpp* start codon and almost half of its coding sequence (the first 288/588 amino acids), FLP/FRT mediated recombination is predicted to yield a null allele. **b**, *dpp<sup>FO-GFP</sup>* heterozygous, *dpp<sup>FO</sup>* heterozygous, and *dpp<sup>FO</sup>* homozygous adult flies. Importantly, *dpp<sup>FO</sup>* homozygous animals have normal adult morphology.



**Extended Data Figure 4 | Validation of an FLP/FRT-mediated conditional allele.** **a**, FRT 5'-loxP and FRT 3' genomic regions are sequenced. **b**, Southern blot analysis of *dpp<sup>FO-GFP</sup>*. Genomic DNAs from *w<sup>1118</sup>* and *dpp<sup>FO-GFP</sup>* are digested by ClaI and are subjected to Southern blot analysis using a GFP probe. **c**, Molecular confirmation of the FLP/FRT-mediated *dpp* FLP-OFF system. As expected, an FLP-OFF product (2,349 bp PCR fragment) is only detected in the *dpp<sup>FO</sup>/dpp<sup>FO</sup>*; *dpp-GAL4/UAS-FLP* lane. Asterisk indicates a non-specific PCR product. **d**, Biochemical evidence of the FLP/FRT-mediated *dpp* FLP-OFF system. *y,w,hs-FLP*; *dpp<sup>FO</sup>/dpp<sup>FO</sup>* larvae are incubated at 37 °C for 30 min at 96 h

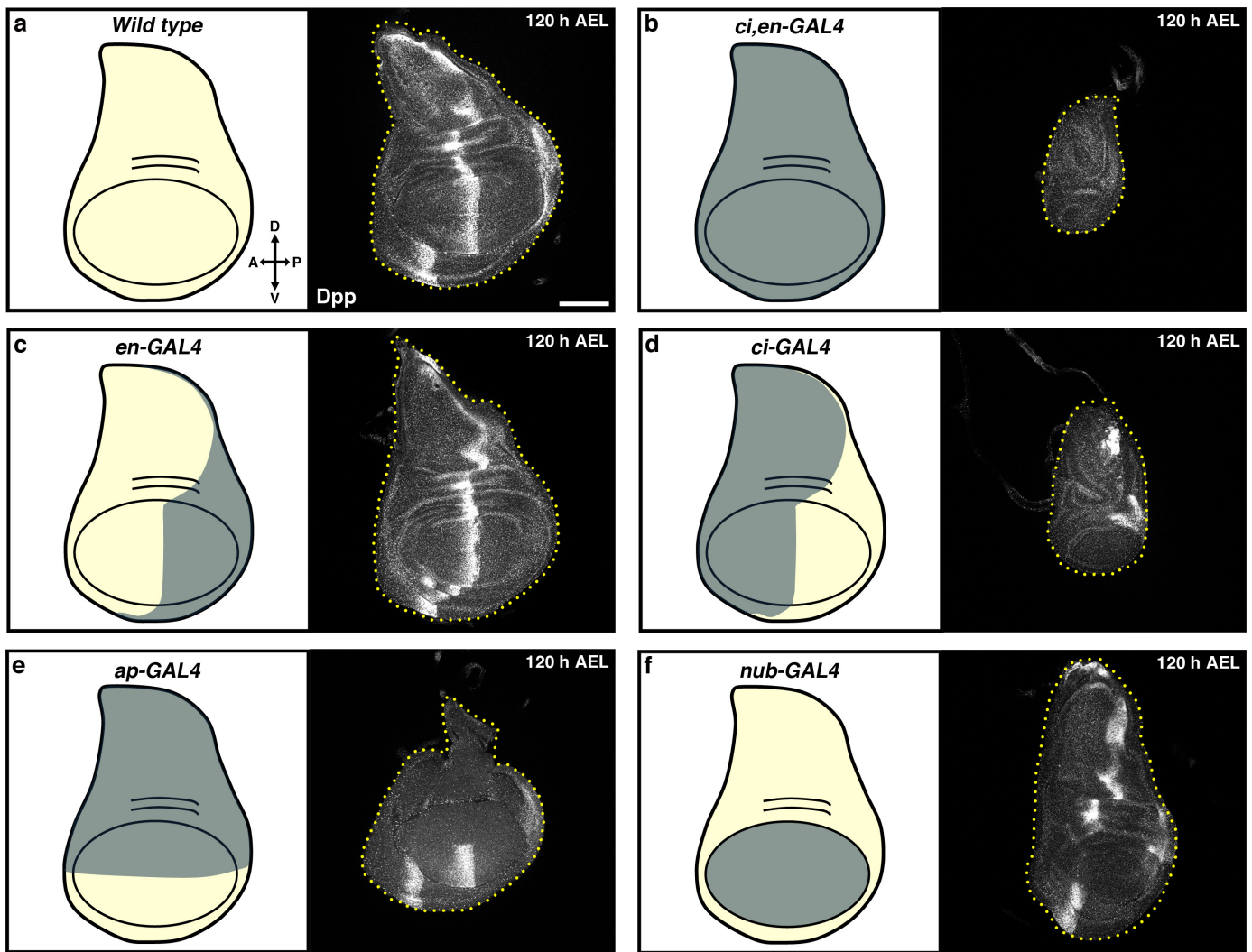
AEL to eliminate Dpp expression. After 24 h, wing discs are homogenized in SDS sample buffer and analysed by western blot analysis with anti-Dpp. Non-specific bands are indicated by an asterisk. Anti-α-tubulin is used as a loading control. **e**, A system to visualize the efficiency of FLP/FRT-mediated recombination. **f–k**, *dpp<sup>FO/+</sup>; dpp-GAL4/UAS-FLP, Act5c(-FRT)lacZ* controls (**f**, **g**, **h**) and *dpp<sup>FO</sup>/dpp<sup>FO</sup>; dpp-GAL4/UAS-FLP, Act5c(-FRT)lacZ* (**i**, **j**, **k**) wing discs are stained with anti-Dpp and β-galactosidase antibodies. The lineage of *dpp-GAL4*-expressing cells is visualized by anti-β-galactosidase staining. Scale bar, 100 μm. Anterior is left.



**Extended Data Figure 5 | p-Mad staining of wing discs lacking *dpp* function in the dorsal compartment. a–d, *dpp<sup>FO</sup>/ap-GAL4,UAS-GFP; UAS-FLP/+* (a, b) and *dpp<sup>FO</sup>/dpp<sup>FO</sup>,ap-GAL4,UAS-GFP; UAS-FLP/+***

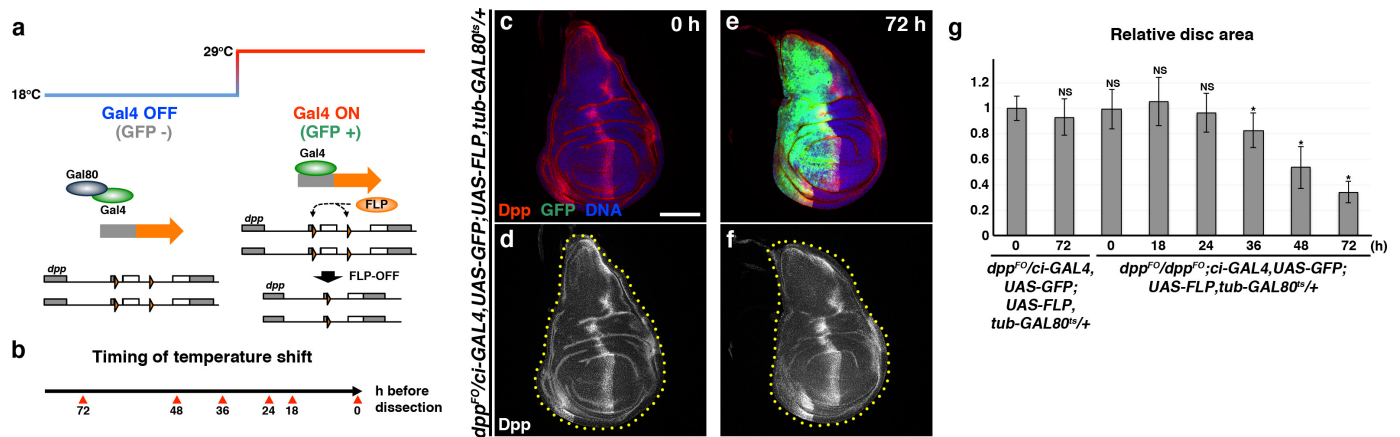
**(c, d)** are dissected and stained with anti-p-Mad antibody. The dorso-ventral boundaries are indicated by green dotted lines. Yellow dotted lines show the disc areas. Scale bar, 50  $\mu\text{m}$ .





**Extended Data Figure 6 | Elimination of Dpp from specific regions of wing discs.** a–f, Anti-Dpp antibody staining of wild-type (a),  $dpp^{FO}$ ,  $ci$ -GAL4,  $en$ -GAL4/ $dpp^{FO}$ ; UAS-FLP/+ (b),  $dpp^{FO}$ ,  $en$ -GAL4/ $dpp^{FO}$ ; UAS-FLP/+ (c),  $dpp^{FO}$ ,  $ci$ -GAL4/ $dpp^{FO}$ ; UAS-FLP/+ (d),  $dpp^{FO}$ ,  $ap$ -GAL4/ $dpp^{FO}$ ; UAS-FLP/+ (e), and  $dpp^{FO}$ ,  $nub$ -GAL4/ $dpp^{FO}$ ; UAS-FLP/+ (f). Gal4-expressing domains are highlighted in grey in each illustration. Wing disc boundaries are shown by dotted lines. Scale bar, 100  $\mu$ m. Anterior is left.

UAS-FLP/+ (e), and  $dpp^{FO}$ ,  $nub$ -GAL4/ $dpp^{FO}$ ; UAS-FLP/+ (f). Gal4-expressing domains are highlighted in grey in each illustration. Wing disc boundaries are shown by dotted lines. Scale bar, 100  $\mu$ m. Anterior is left.



**Extended Data Figure 7 | Spatiotemporal Dpp removal from the anterior region of wing discs.** **a**, A strategy for temporal Dpp elimination from the anterior compartment of wing discs using the *GAL80<sup>ts</sup>* system. At 18 °C, Gal4 activity is blocked by Gal80. When flies are kept at 29 °C (non-permissive temperature for *GAL80<sup>ts</sup>*), Gal4 induces expression of FLP and GFP. **b**, Timing of temperature shift. Larvae are reared at 18 °C, and are transferred to 29 °C at the indicated time points before dissection. **c–f**, *dpp<sup>FO</sup>/ci-GAL4, UAS-GFP; UAS-FLP, tub-GAL80<sup>ts</sup>/+* controls are

stained with anti-Dpp. Gal4 activity is monitored by GFP expression. Scale bar, 100  $\mu$ m. **g**, Size comparison between wing discs: *dpp<sup>FO</sup>/ci-GAL4, UAS-GFP; UAS-FLP, tub-GAL80<sup>ts</sup>/+* (0 ( $n = 21$ ) and 72 h ( $n = 40$ ) before dissection) and *dpp<sup>FO</sup>/dpp<sup>FO</sup>; ci-GAL4, UAS-GFP; UAS-FLP, tub-GAL80<sup>ts</sup>/+* (0 ( $n = 36$ ), 18 ( $n = 33$ ), 24 ( $n = 40$ ), 36 ( $n = 33$ ), 48 ( $n = 80$ ) and 72 h ( $n = 43$ ) before dissection). Mean  $\pm$  s.d. \* $P < 0.001$ , not significant (NS), two-sided Student's  $t$ -test.

# Diversion of aspartate in ASS1-deficient tumours fosters *de novo* pyrimidine synthesis

Shiran Rabinovich<sup>1\*</sup>, Lital Adler<sup>1\*</sup>, Keren Yizhak<sup>2</sup>, Alona Sarver<sup>1</sup>, Alon Silberman<sup>1</sup>, Shani Agron<sup>1</sup>, Noa Stettner<sup>1</sup>, Qin Sun<sup>3</sup>, Alexander Brandis<sup>4</sup>, Daniel Helbling<sup>5</sup>, Stanley Korman<sup>6</sup>, Shalev Itzkovitz<sup>7</sup>, David Dimmock<sup>5</sup>, Igor Ulitsky<sup>1</sup>, Sandesh C. S. Nagamani<sup>3,8</sup>, Eytan Ruppin<sup>2,9,10</sup> & Ayelet Erez<sup>1</sup>

Cancer cells hijack and remodel existing metabolic pathways for their benefit. Argininosuccinate synthase (ASS1) is a urea cycle enzyme that is essential in the conversion of nitrogen from ammonia and aspartate to urea. A decrease in nitrogen flux through ASS1 in the liver causes the urea cycle disorder citrullinaemia<sup>1</sup>. In contrast to the well-studied consequences of loss of ASS1 activity on ureagenesis, the purpose of its somatic silencing in multiple cancers is largely unknown<sup>2</sup>. Here we show that decreased activity of ASS1 in cancers supports proliferation by facilitating pyrimidine synthesis via CAD (carbamoyl-phosphate synthase 2, aspartate transcarbamylase, and dihydroorotase complex) activation. Our studies were initiated by delineating the consequences of loss of ASS1 activity in humans with two types of citrullinaemia. We find that in citrullinaemia type I (CTLN I), which is caused by deficiency of ASS1, there is increased pyrimidine synthesis and proliferation compared with citrullinaemia type II (CTLN II), in which there is decreased substrate availability for ASS1 caused by deficiency of the aspartate transporter citrin. Building on these results, we demonstrate that ASS1 deficiency in cancer increases cytosolic aspartate levels, which increases CAD activation by upregulating its substrate availability and by increasing its phosphorylation by S6K1 through the mammalian target of rapamycin (mTOR) pathway. Decreasing CAD activity by blocking citrin, the mTOR signalling, or pyrimidine synthesis decreases proliferation and thus may serve as a therapeutic strategy in multiple cancers where ASS1 is downregulated. Our results demonstrate that ASS1 downregulation is a novel mechanism supporting cancerous proliferation, and they provide a metabolic link between the urea cycle enzymes and pyrimidine synthesis.

In contrast to the well-delineated biochemical and clinical consequences of loss-of-function germline mutations in ASS1, which have not been reported to include cancer, studies have shown a correlation between somatic deficiency of ASS1 in cancer and poor prognosis, for which the mechanism remains obscure<sup>2,3</sup>. Outside the liver, ASS1 is expressed in most tissues where it catalyses the penultimate step in the synthesis of arginine. Argininosuccinate lyase (ASL), the enzyme downstream of ASS1, is directly responsible for arginine synthesis<sup>4</sup> (Fig. 1a). A well-established sequel of ASS1 and/or ASL deficiency is arginine auxotrophy<sup>5</sup>; thus, arginine-catabolizing enzymes have been used as therapy in ASS1-depleted tumours with limited benefit, especially in melanoma, wherein the cancer cells develop resistance by re-expressing ASS1 within days<sup>3</sup>. Since there are cancers in which both these genes are epigenetically silenced<sup>6</sup>, ASS1 deficiency in cancers might have an arginine-independent effect, which may be related to its substrate, aspartate (Fig. 1a).

In the cytosol, aspartate serves as a substrate both for ASS1 and for the enzymatic complex CAD. We thus hypothesized that decreased ASS1 activity might enhance aspartate availability for CAD for the synthesis of pyrimidine nucleotides to promote proliferation (Fig. 1a). If correct, deficiency in the mitochondrial aspartate transporter, citrin, would be expected to decrease aspartate availability both for ASS1 and for CAD and hence restrict proliferation (Fig. 1a).

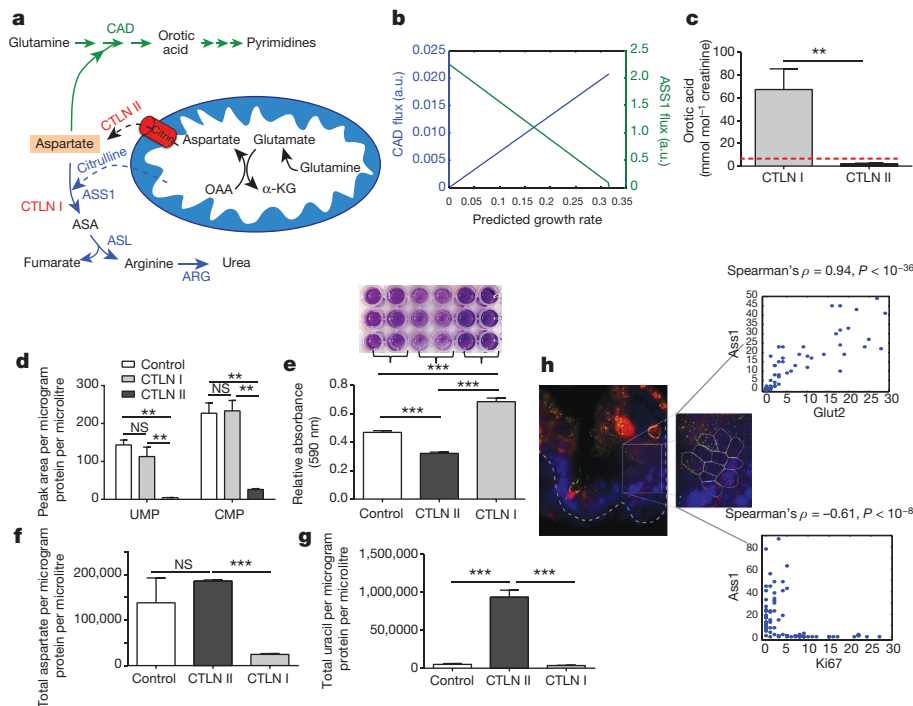
We first assessed the correlation between ASS1 levels and proliferation in non-cancerous states. A generic stoichiometric model of human metabolism<sup>7</sup> predicted that inactivation of ASS1 is significantly associated with an increase in growth rate, and is additionally predicted to increase flux through the reaction catalysed by CAD (Fig. 1b). Thus, we expected subjects with ASS1 deficiency (CTLN I) to have increased synthesis of pyrimidines owing to increased utilization of aspartate by CAD, compared with those with CTLN II in whom aspartate availability to CAD is decreased (Fig. 1a). Indeed, urinary levels of orotic acid, a product reflecting the synthetic activity of CAD, were significantly elevated in human subjects with CTLN I compared with the normative values from a control population and with subjects with CTLN II (Fig. 1a, c). Moreover, we found that CTLN I fibroblasts have increased synthesis of pyrimidines and proliferation compared with CTLN II cells (Fig. 1d, e). Using [<sup>15</sup>N<sub>5</sub>]α-glutamine we further showed that CTLN I cells generate more total as well as labelled M + 1 aspartate and M + 1 uracil, compared with control and CTLN II fibroblasts (Fig. 1f, g and Extended Data Fig. 1a–c). Hence, there is a specific decrease in aspartate transport from the mitochondria in CTLN II, leading to reduced aspartate availability for pyrimidine synthesis and restricting proliferation. Interestingly, growth restriction has been reported in humans with CTLN II<sup>8</sup> but no growth aberrancies have been reported in CTLN I, further providing a clinical human context to the findings and suggesting that, in physiological proliferation, aspartate deficiency has more severe clinical consequences than its enrichment. To corroborate our results in another model system, we analysed *Ass1* messenger RNA (mRNA) levels in wild-type newborn mouse intestines, which express high levels of *Ass1* and contain proliferating and differentiated cells in the crypts and villi, respectively<sup>9</sup>. We found a significant correlation between the levels of *Ass1* and *Glut2*, a mature enterocyte marker in the differentiated enterocytes in the villi, whereas a significant inverse correlation was observed between *Ass1* and *Ki67*, a marker of proliferation, in the proliferating cells in the crypts (Fig. 1h). Thus, ASS1 inactivation has an important role in proliferation of non-cancerous cells, by increasing aspartate availability for pyrimidine synthesis by CAD.

We next evaluated whether this mechanism could be the reason for the downregulation of ASS1 in cancer. According to the well-established ‘Warburg effect’, different metabolites are diverted from

<sup>1</sup>Department of Biological Regulation, Weizmann Institute of Science, Rehovot 7610001, Israel. <sup>2</sup>The Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel. <sup>3</sup>Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030, USA. <sup>4</sup>Biological Services, Weizmann Institute of Science, Rehovot 69978, Israel. <sup>5</sup>Human and Molecular Genetic and Biochemistry Center, Medical College Wisconsin, Milwaukee, Wisconsin 53226, USA. <sup>6</sup>Genetic and Metabolic Center, Hadassah Medical Center, Jerusalem 91120, Israel. <sup>7</sup>Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot 69978, Israel. <sup>8</sup>Texas Children's Hospital, Houston, Texas 77030, USA. <sup>9</sup>The Sackler School of Medicine, Tel Aviv University, Tel Aviv 69978, Israel. <sup>10</sup>Center for Bioinformatics and Computational Biology & Department of Computer Science, University of Maryland, College Park, Maryland 20742, USA.

\*These authors contributed equally to this work.



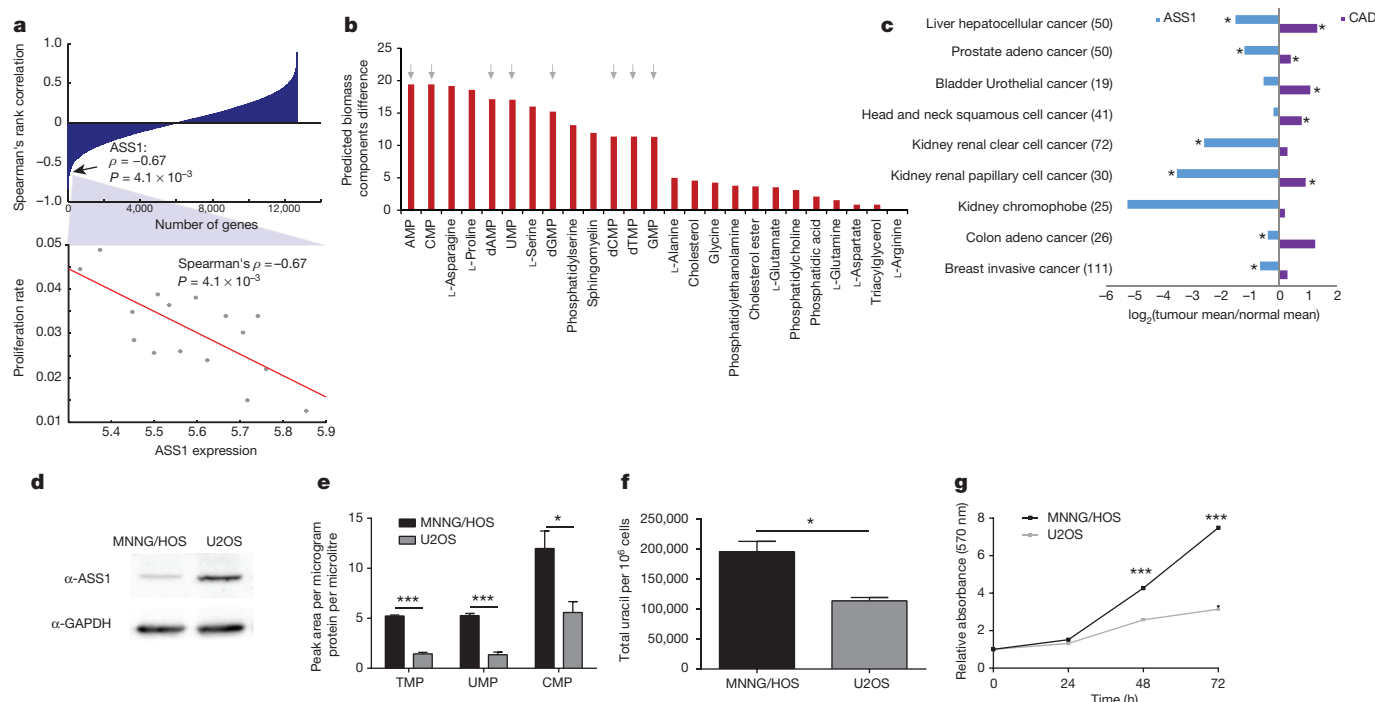


**Figure 1 | ASS1 inactivation correlates with non-cancerous proliferation.** **a**, Illustration of the metabolic flux involved in nitrogen contributions to nucleic acid synthesis. The aspartate nitrogen can be used for synthesis of pyrimidines (green path) or urea (blue path). In ASS1 deficiency (CTLN I) there is a potential diversion of the aspartate towards pyrimidine whereas in citrin deficiency (CTLN II) aspartate is not transported across the mitochondria. ASA, argininosuccinate; ARG, arginase. **b**, Prediction by the generic human model; decreasing ASS1 activation (green line) results in an increase in the cellular growth rate and in the flux through the CAD reaction (blue line); a.u., arbitrary units. **c**, Urinary orotic acid levels are elevated significantly in patients with CTLN I compared with normative values in control subjects (0.3–2.8 mmol mol<sup>-1</sup> creatinine, depicted by the red dashed line), and with those with CTLN II.  $^{**}P < 0.005$  using log-transformed data for *t*-test analysis ( $n = 5$  with CTLN I and  $n = 4$  with CTLN II). **d–g**, These experiments were repeated twice with pooled cells from two patients with CTLN II, from one patient with CTLN I and from three control subjects. Statistical analysis used one-way analysis of variance (ANOVA). Error bars represent standard error. **d**, Significantly lower levels of pyrimidines in

their 'routine pathways' for the synthesis of biological molecules that are essential for cell division and growth. We hence conducted an analysis of *ASS1* expression data in cancer cell lines from the NCI-60 collection and found a significant inverse correlation between *ASS1* expression levels and the reported doubling time of the cancerous cells (Fig. 2a). To further test whether this correlation is explicable by diversion of aspartate flux, we used our modelling program and predicted that, with *ASS1* inactivation, there is an accompanying significant increase in aspartate flux through the relevant metabolic reactions for nucleic acid synthesis (Extended Data Table 1). In contrast, modelling the inactivation of *ASL* predicted an endogenous arginine depletion that does not directly affect the flux towards nucleic-acid synthesis (data not shown). Furthermore, analysis of The Cancer Genome Atlas database for *ASL* and *ASS1* expression shows that these genes can both be downregulated in the same cancers, suggesting that they are not mutually exclusive (Extended Data Fig. 1d). Thus, *ASS1* silencing in cancerous proliferation might have an arginine-independent effect that is related to nucleotide synthesis.

Using specific metabolic models tailored for each of the NCI-60 cell lines<sup>10</sup>, we further predicted that 8 out of the 13 metabolites computationally shown to be increased with *ASS1* inactivation were nucleic acids (Fig. 2b and Extended Data Fig. 1e). Additionally, specific analysis

of the TCGA database of tumours where *ASS1* expression is downregulated showed a significant upregulation in the expression of *CAD*, compared with the paired normal tissue (Fig. 2c). We further confirmed the inverse upregulation in the expression of *CAD* versus *ASS1* at the mRNA level in the NCI-60 cancer cell-line database as well as in independent databases for patients with osteosarcoma<sup>11</sup> and melanoma<sup>12</sup>; we found that downregulation of *ASS1* and upregulation of *CAD* are in concordance with cancerous phenotype (Extended Data Fig. 1f, g). In addition, we demonstrated the inverse expression levels between *ASS1* and *CAD* at the protein level using osteosarcoma and melanoma cell lines that differ in their expression pattern of *ASS1* (Extended Data Figs 1h and 2a). To validate these modelling and global informatics analyses with experimental evidence, we studied osteosarcoma cell lines in which *ASS1* was either deficient (MNNG/HOS) or present (U2OS) (Fig. 2d and Supplementary Fig. 1). Metabolic analyses confirmed that cells deficient in *ASS1* had an increase in pyrimidine levels, an increase in the level of uracil as well as a significantly increased proliferation rate (Fig. 2e–g) compared with osteosarcoma cells having higher levels of *ASS1*. We additionally verified these results in melanoma cell lines that differed in their level of *ASS1* expression (Extended Data Fig. 2b–d). To definitively demonstrate the direct correlation between *ASS1* expression and proliferation and to differentiate it from other metabolic



**Figure 2 | ASS1-deficient tumours have increased proliferation rate and pyrimidine levels.** **a**, Top: ASS1 is ranked within the top 24 genes out of 14,000, that show a significant inverse correlation with proliferation rate among the 16 NCI-60 cancer cell lines in which ASS1 is downregulated; Spearman's rank correlation was calculated between the expression of each gene and its associated proliferation rates. Bottom: magnified view of the correlation between proliferation rate and ASS1 expression levels. **b**, Predicted differences in the production rate of biomass components after the inactivation of ASS1. The production of nucleic acids marked in arrows, is predicted to have a large increase after ASS1 inactivation in most of the NCI-60 cell lines. The figure represents the results obtained using the LOX IMVI cell line model; however, the same results were seen for all NCI-60 cell lines as well as in using the human generic model.

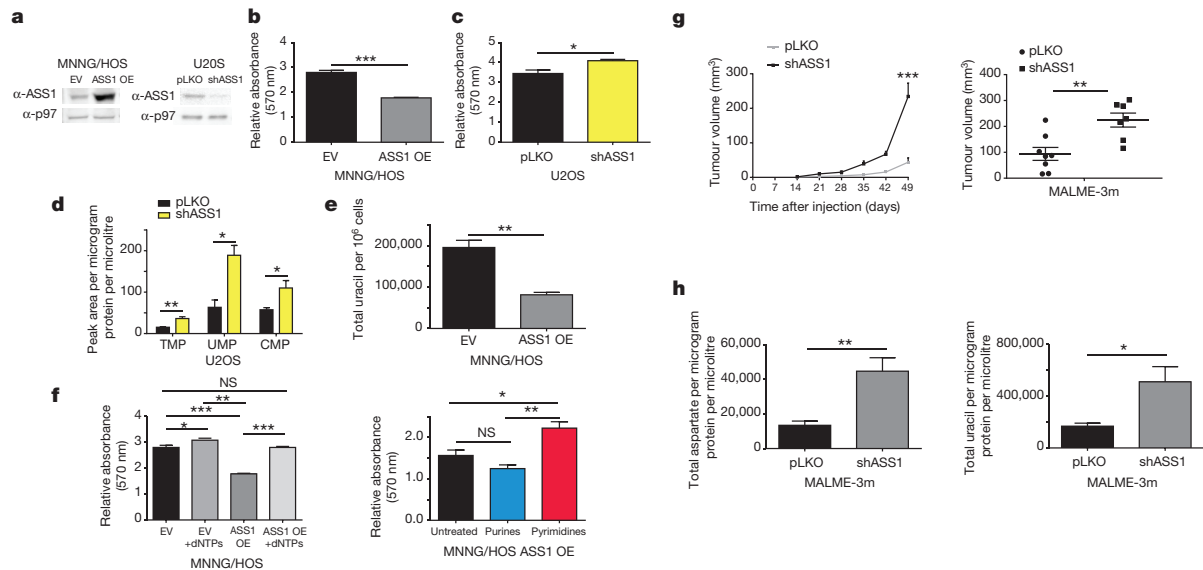
changes that occur in cancer cells, we overexpressed ASS1 in MNNG/HOS, and knocked it down in U2OS cells (Fig. 3a). Our results clearly show that changes in ASS1 levels inversely alter the proliferation rate and pyrimidine synthesis in these cells (Fig. 3b–e and Extended Data Fig. 3a–f). If the major determinant by which ASS1 overexpression decreases proliferation is through diverting aspartate metabolism away from pyrimidine synthesis, supplementation with nucleic acids should restore proliferation. Indeed, supplementing the media with nucleic acids, specifically with pyrimidines, significantly restores the proliferation of ASS1-overexpressing cells to a similar level as the parental cell line (Fig. 3f and Extended Data Fig. 2e–j). Thus, in two distinct forms of cancer, changes in ASS1 expression levels directly affect aspartate utilization for pyrimidine synthesis and proliferation. Importantly, similar results were obtained *in vivo*; mice injected with melanoma cells knocked down for ASS1 developed tumours that grew more rapidly and had higher levels of total and M + 1-labelled aspartate and uracil than the parental tumour cells that expressed the empty vector (Fig. 3g, h and Extended Data Fig. 3g).

An expected synergistic way to increase aspartate delivery for pyrimidine synthesis would be by upregulation of citrin. Analysis of the TCGA database showed that in tissues that normally do not express citrin (also known as SLC25A13) at high levels<sup>13</sup>, there is a significantly elevated expression in the cancerous state (Extended Data Fig. 4b). In addition, in the liver where citrin is strongly expressed, a recent publication of ASS1 expression in hepatocellular carcinoma showed that downregulation of ASS1 is associated with a more malignant cancerous phenotype<sup>14</sup>. These results, together with our study of primary human fibroblast cells (Fig. 1c–g), imply that proliferation

The models used are based on a series of simplifying assumptions as described by us previously in detail<sup>10</sup>. **c**, Analysis of the TCGA database of matched tumour–normal tissue pairs showing that CAD expression is elevated significantly in tumours with ASS1 downregulation compared with normal tissue ( $*P < 0.05$ ). **d**, Immunoblot of osteosarcoma cell lines showing decreased expression of ASS1 compared with the loading control GAPDH in MNNG/HOS compared with U2OS. **e**, Osteosarcoma cells with ASS1 downregulation have a significant increase in pyrimidine levels as measured by LC–MS ( $n \geq 3$ ;  $*P < 0.05$ ;  $***P < 0.0005$ ). Error bars are standard error. **f**, Osteosarcoma cells with ASS1 downregulation have a significant increase in uracil ( $n = 3$ ;  $*P < 0.05$ ). Error bars are standard error. **g**, Osteosarcoma cells with ASS1 downregulation have a significant increase in proliferation as measured by MTT assay ( $***P < 0.0005$ ).

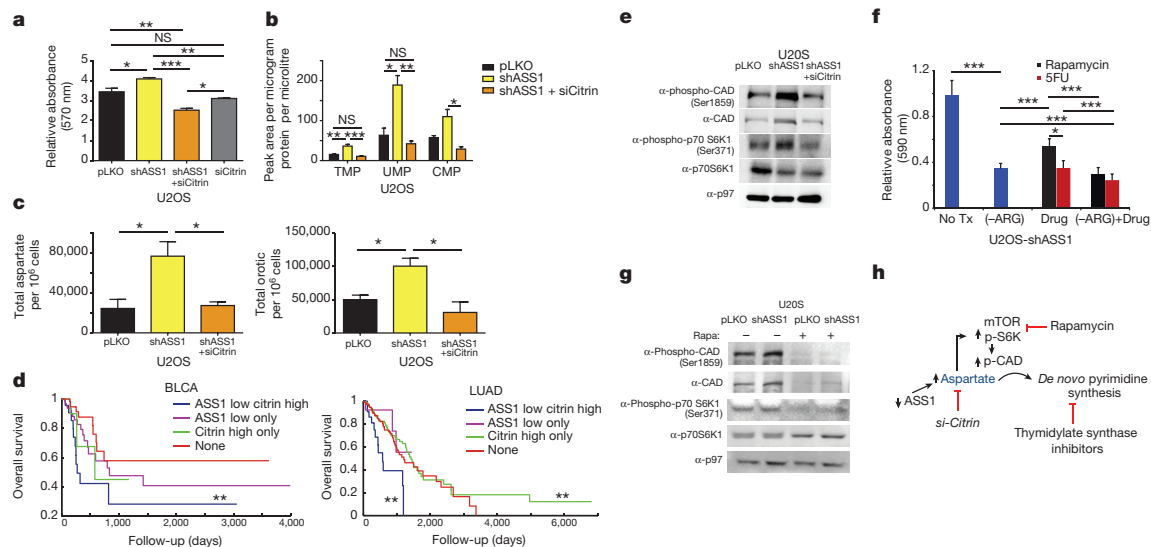
induced by loss of ASS1 in tumours might be counteracted by inhibiting citrin. Indeed, *si-citrin* in U2OS decreases proliferation significantly when ASS1 levels are reduced (Fig. 4a). Use of *si-citrin* also decreases pyrimidine levels as well as total and labelled levels of M + 1 aspartate and of M + 1 orotic acid (Fig. 4b, c and Extended Data Fig. 4c, d). As citrin is part of the malate–aspartate shuttle, its deficiency is expected to affect several aspects in cell survival and growth. Our results indicate that citrin function in transferring mitochondrial-derived aspartate is important for supplying substrate for pyrimidine synthesis, especially in cancers with ASS1 downregulation. These findings are therapeutically relevant as survival analysis of several cancers in the TCGA database reveals that cancers with decreased ASS1 expression and high *citrin* levels have a trend for significantly worse prognosis (Fig. 4d, Extended Data Fig. 4e and Extended Data Table 2).

The use of citrin-derived aspartate by CAD requires CAD activation. Recently, CAD was shown to be activated by ribosomal protein S6 kinase (S6K1), regulated by the mTOR pathway<sup>15</sup>. When ASS1 expression in cancer cells is decreased, we find increased phosphorylation of S6K1 and CAD that is reduced by *si-citrin*, implying that aspartate levels are important in regulating the mTOR pathway activation (Fig. 4e and Extended Data Fig. 4f). In addition, we show a significant increase in the location proximity between CAD and citrin after ASS1 downregulation (Extended Data Fig. 4g). Thus, aspartate regulates pyrimidine levels by regulating CAD's substrate availability, protein localization and activity. Consistent with this, we see a decrease in proliferation when ASS1-deficient cells are treated either with the mTOR inhibitor rapamycin or with thymidylate synthase inhibitor 5-fluorouracil (5FU) (Fig. 4f). Importantly, rapamycin treatment is accompanied by a



**Figure 3 | ASS1 expression levels in cancer determine aspartate availability for pyrimidine synthesis.** **a**, Immunoblots of osteosarcoma cells after transduction with either ASS1 overexpression construct (left) or with ASS1-shRNA (right). **b**, MTT proliferation assay in osteosarcoma cells showing a significant decrease in proliferation after ASS1 overexpression ( $***P < 0.0005$ ). **c**, MTT proliferation assay in osteosarcoma cells showing a significant increase in proliferation after transduction with *shASS1* ( $*P < 0.05$ ). The proliferation values are shown for day 3, after normalizing the data for the reading after cell adherence ( $n \geq 3$ ). **d**, LC-MS measurements of pyrimidine levels showing a significant increase after the use of ASS1-shRNA in osteosarcoma cells ( $n \geq 3$ ). **e**, Total uracil is decreased significantly in osteosarcoma cells with ASS1 overexpression ( $n \geq 3$ ;  $**P = 0.005$ ). **f**, Left: dNTP supplementation rescues proliferation after ASS1 overexpression in osteosarcoma cells. Right: pyrimidines significantly rescue

proliferation in ASS1-overexpressing MNNG/HOS cells ( $n \geq 3$ ;  $*P < 0.05$ ,  $**P < 0.005$ ,  $***P < 0.0005$ ). **g**, **h**, Ten million MALME-3m melanoma cells transduced with either pLKO empty vector or with *shASS1* were injected subcutaneously to immune-deficient mice. The experiment was repeated three times. After euthanasia, the tumours were removed, measured and incubated with labelled [ $^{15}\text{N}$ ]α-glutamine for 6 h. Two weeks after injection, the group injected with melanoma cells with *shASS1* developed tumours that grew more rapidly in size (**g**, left) and were hence significantly larger when removed (**g**, right). The experiment was repeated three times with similar results ( $**P < 0.005$ ,  $***P < 0.0005$ ). **h**, Tumours with *shASS1* had higher levels of total aspartate (left) and total uracil (right) than those expressing the empty vector. Statistical analysis used repeated-measurements ANOVA ( $n = 15$ ;  $*P < 0.05$ ;  $**P < 0.005$ ). All error bars are standard error.



**Figure 4 | Decreasing CAD activation reduces proliferation in ASS1-deficient cancers.** **a**, MTT assay showing that decreasing citrin levels significantly decreases proliferation in U2OS osteosarcoma, even after a significant proliferation increase is accomplished by ASS1 downregulation ( $n \geq 3$ ;  $*P < 0.05$ ;  $**P < 0.005$ ;  $***P < 0.0005$ ). **b**, Decreasing citrin levels decreases pyrimidine levels in U2OS cells with ASS1 downregulation ( $n \geq 3$ ;  $*P < 0.05$ ). **c**, GC-MS measurements showing that U2OS with *shASS1* has a significant increase in total aspartate (left), as well as in total orotic acid (right), which are reversed when transfected with *si-citrin* ( $n \geq 3$ ; ANOVA, Tukey's honest significant difference test,  $*P < 0.05$ ). Error bars are standard error. **d**, Kaplan-Meier survival analysis for two different cancer types (BLCA, bladder cancer; LUAD, lung adenocarcinoma), both showing significantly poor survival for cancers with low ASS1 and high citrin expression levels ( $n \geq 3$ ;  $**\log \text{rank } P \leq 0.005$ ). **e**, Immunoblot of osteosarcoma cells for the mTOR pathway downstream effectors S6K1 and CAD showing increased

phosphorylation after *shASS1* that is reversed when cells are transfected with *si-citrin*. **f**, Quantification graph of crystal violet staining of osteosarcoma cells transduced with *shASS1* after drug treatments (Tukey's post-hoc;  $P < 0.0005$ ). All treatments were significant compared with no treatment. In addition, the results show a significant beneficial effect of decreased proliferation in response to treatment with either mTOR or pyrimidine synthesis inhibitors (rapamycin or 5FU respectively), with 5FU being more beneficial than rapamycin ( $P < 0.0005$ ). Of note, 5FU had a significant additive beneficial effect when added as treatment to arginine-depleted medium (Tukey's post-hoc test;  $P < 0.0005$ ). Cells were grown in normal medium, in medium depleted of arginine, in complete medium with either rapamycin or 5FU, and in arginine-depleted medium together with either rapamycin or 5FU ( $n = 9$ ;  $***P < 0.0005$ ). **g**, Western blot showing decreased activation of the mTOR proteins after rapamycin treatment. **h**, Schematic presentation for potential interventions in pyrimidine synthesis in ASS1-deficient tumours.



decrease in CAD phosphorylation (Fig. 4g). Hence, targeting aspartate transport could be an additional therapeutic option in cancers with ASS1 silencing, especially in those that develop resistance to arginine-depleting agents (Fig. 4h).

In summary, our studies demonstrate that ASS1, a urea cycle enzyme, facilitates pyrimidine synthesis in cancerous proliferation by activating CAD, through regulation of aspartate levels. There are several clinical trials in patients with ASS1-deficient hepatocellular carcinoma and mesothelioma, which combine arginine-depleting agents with thymidylate synthase inhibitors such as capecitabine and pemetrexed (<http://clinicaltrials.gov>, NCT02089633, NCT02029690). We believe our study provides the rationale for such therapeutic modalities and hence has direct translational relevance.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 17 November 2014; accepted 27 August 2015.**

**Published online 11 November 2015.**

- Dimmock, D. *et al.* Citrin deficiency, a perplexing global disorder. *Mol. Genet. Metab.* **96**, 44–49 (2009).
- Delage, B. *et al.* Arginine deprivation and argininosuccinate synthetase expression in the treatment of cancer. *Int. J. Cancer* **126**, 2762–2772 (2010).
- Long, Y. *et al.* Arginine deiminase resistance in melanoma cells is associated with metabolic reprogramming, glucose dependence, and glutamine addiction. *Mol. Cancer Ther.* **12**, 2581–2590 (2013).
- Morris, S. M., Jr. Recent advances in arginine metabolism: roles and regulation of the arginases. *Br. J. Pharmacol.* **157**, 922–930 (2009).
- Wheatley, D. N. Controlling cancer by restricting arginine availability—arginine-catabolizing enzymes as anticancer agents. *Anticancer Drugs* **15**, 825–833 (2004).
- Syed, N. *et al.* Epigenetic status of argininosuccinate synthetase and argininosuccinate lyase modulates autophagy and cell death in glioblastoma. *Cell Death Dis.* **4**, e458 (2013).
- Duarte, N. C. *et al.* Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl Acad. Sci. USA* **104**, 1777–1782 (2007).
- Kobayashi, K., Saheki, T. & Song, Y. Z. Citrin deficiency. *GeneReviews* <http://www.ncbi.nlm.nih.gov/books/NBK1181/> (2005).
- Marion, V. *et al.* Hepatic adaptation compensates inactivation of intestinal arginine biosynthesis in suckling mice. *PLoS One* **8**, e67021 (2013).
- Yizhak, K. *et al.* Phenotype-based cell-specific metabolic modeling reveals metabolic liabilities of cancer. *eLife* **3**, e03641 (2014).
- Kuijjer, M. L. *et al.* IGF1R signaling as potential target for treatment of high-grade osteosarcoma. *BMC Cancer* **13**, 245 (2013).
- Kabbarah, O. *et al.* Integrative genome comparison of primary and metastatic melanomas. *PLoS One* **5**, e10770 (2010).
- del Arco, A. *et al.* Expression of the aspartate/glutamate mitochondrial carriers aralar1 and citrin during development and in adult rat tissues. *Eur. J. Biochem.* **269**, 3313–3320 (2002).
- Tan, G. S. *et al.* Novel proteomic biomarker panel for prediction of aggressive metastatic hepatocellular carcinoma relapse in surgically resectable patients. *J. Proteome Res.* **13**, 4833–4846 (2014).
- Ben-Sahra, I., Howell, J. J., Asara, J. M. & Manning, B. D. Stimulation of de novo pyrimidine synthesis by growth signaling through mTOR and S6K1. *Science* **339**, 1323–1328 (2013).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank A. Gross and B. Lee for discussions.

We acknowledge and thank the Weizmann Institute for providing financial and infrastructural support and the Baylor College of Medicine Biochemical Laboratory. We appreciate the statistical analysis by R. Rotkopf and the technical contributions of A. Tishbee, T. Kaufman, D. Laufer and I. Rogachev. A.E. is the incumbent of the Leah Omenn Career Development Chair and is supported by research grants from the European Research Program (CIG618113, ERC614204), the Israel Science Foundation (1343/13; 1952/13) and a Minerva grant award (711730). A.E. received additional support from the Adelis Foundation, the Henry S. and Anne S. Reich Research Fund, the Dukler Fund for Cancer Research, the Paul Sparr Foundation, the Saul and Theresa Esman Foundation, from Joseph Piko Baruch and from the estate of Fannie Sherr. L.A. was supported by a postdoctoral fellowship from Teva. The research of K.Y. and E.R. has been supported by grants from the Israeli Science Foundation (E.R.), the Israeli Cancer Research Fund (E.R.), the Israeli Center of Excellence (I-CORE) Program of the Planning and Budgeting Committee and Israel Science Foundation Grant No. 41/11 (E.R.). A.Sa. is supported by the Israel Cancer Research Foundation and S.N.S.C. is supported by a Baylor College of Medicine Intellectual and Developmental Disabilities Research Center Grant (number 1 U54 HD083092) from the Eunice Kennedy Shriver National Institute of Child Health & Human Development, and by the Doris Duke Charitable Foundation (DDCF 2013095). I.U. was supported by a grant from the Rising Tide Foundation and by a research grant from The Abramson Family Center for Young Scientists.

**Author Contributions** S.R. performed most of the experiments described in the manuscript; L.A. set up the system that allowed us to test our hypothesis; K.Y. and E.R. designed the modelling analysis. K.Y. from the laboratory of E.R. performed the modelling analysis; A.Sa. performed *in vitro* studies with the patients' cells and helped establish the *in vivo* models; A.Si. executed the metabolic analysis by gas chromatography–mass spectrometry (GC–MS); S.A. helped with the *si-citrin* experiments; N.S. performed the *in vivo* experiments; Q.S. analysed levels of orotic acid in the patients' urine; A.B. performed the pyrimidine analysis by liquid chromatography–mass spectrometry (LC–MS); D.H., S.K. and D.D. provided data and primary cell lines from patients with CTLN II; S.I. and S.A. performed the fluorescence *in situ* hybridization (FISH) experiments of the mouse intestine; I.U. analysed the TCGA database; S.C.S.N. provided primary cells of patients with CTLN I as well as help writing the manuscript; A.E. was the leading principal investigator who initiated and directed the study and co-wrote the paper with inputs from all authors. K.Y., A.Si., S.H. and A.Sa. contributed equally to this work.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.E. (ayelet.erez@weizmann.ac.il).

## METHODS

Unless mentioned otherwise, the experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

**Measurements in human subjects.** The fibroblast studies were performed on anonymized cells devoid of all identifiers. The data analysis involving urine orotic acid levels were performed under a protocol approved by the Institutional Review Board of Baylor College of Medicine. Urine samples were prepared by mixing 200  $\mu$ l of with isotopic internal standard [ $^{15}\text{N}_2$ ]orotic acid (Cambridge Isotope Laboratories). Orotic acid and orotidine were assayed on a Micromass Quattro mass spectrometer (Waters). HPLC was performed on a Waters ODS-AQ analytical column (150 mm  $\times$  2.0 mm internal diameter 5  $\mu$ m bead size). Mobile phase was isocratic 0.05 M ammonium formate (pH 4.0). The MS–MS system was set at a flow rate of 0.2 ml min $^{-1}$ . The mass spectrometer was operated in electrospray ionization negative multiple-reaction monitoring mode. Nitrogen was used as nebulizer gas at a flow rate of 60–90 l h $^{-1}$  and desolvation gas 500 l h $^{-1}$ . Other optimized mass spectrometer parameters were cone voltage 15 V, capillary 3,250 V and collision voltage 10 V.

**Genome-scale metabolic modelling.** A metabolic network consisting of  $m$  metabolites and  $n$  reactions can be represented by a stoichiometric matrix  $S$ , where the entry  $S_{ij}$  represents the stoichiometric coefficient of metabolite  $i$  in reaction  $j$ <sup>17</sup>. A constraint-based model imposes mass balance, directionality and flux capacity constraints on the space of possible fluxes in the metabolic network's reactions through a set of linear equations:

$$S \times v = 0 \quad (1)$$

$$v_{\min} \leq v \leq v_{\max} \quad (2)$$

where  $v$  stands for the flux vector for all of the reactions in the model (that is, the flux distribution). The exchange of metabolites with the environment is represented as a set of exchange (transport) reactions, enabling a pre-defined set of metabolites to be either taken up or secreted from the growth media. The steady-state assumption represented in equation (1) constrains the production rate of each metabolite to be equal to its consumption rate. Enzymatic directionality and flux capacity constraints define lower and upper bounds on the fluxes and are embedded in equation (2). In the following, flux vectors satisfying these conditions will be referred to as feasible steady-state flux distributions. The analyses were performed under the Roswell Park Memorial Institute Medium (RPMI)-1640m. We used the biomass function introduced in ref. 16.

**Predicting growth rate, metabolite production and flux distribution through metabolic modelling.** To determine the relation between ASS1 activity, CAD activity and growth rate, we used a generic human model and simulated the inactivation and activation of the reaction catalysed by ASS1. The inactivation was simulated by constraining the flux through the ASS1 reaction to zero, while the activation was simulated by enforcing increased positive flux through the ASS1 reaction up to the maximal possible flux, as computed via flux variability analysis<sup>17</sup>. At each such point, the maximal growth rate was computed via flux balance analysis<sup>17</sup>. Additionally, we estimated the flux through the reaction catalysed by CAD under maximal growth rate on the basis of 1,000 different feasible flux samples<sup>18</sup>.

We next used genome-scale metabolic models for each of the NCI-60 cancer cell lines on the basis of their gene expression measurements<sup>10</sup>. In each cell-line model, we performed the following analyses. (1) We computed the production of each biomass component under both the inactivation and maximal activation of ASS1, as described above. The difference between the predicted production rates of each biomass component in the two states was then computed on the basis of the results of this optimization problem. (2) Similarly, we examined the flux change of each reaction under maximal biomass production in both the inactivation and activation states, as described above. In each of these states, we sampled the solution space and obtained 1,000 feasible flux distributions<sup>18</sup>. Focusing on the reactions associated with aspartate and glutamine, we computed the fold-change in flux rate together with its significance level. The latter was computed via a two-sided Wilcoxon rank-sum test. The largest fold-change among these reactions was predicted for the reactions catalysed by the CAD enzyme.

**TCGA data analysis.** For each tumour, normalized gene expression levels measured using RSEM<sup>19</sup> were obtained from the RNASeqV2 data sets at the TCGA portal (<https://tcga-data.nci.nih.gov/tcga/>). Only matched tumour–normal pairs were used. For each tumour type, we computed the mean expression levels in the tumour and normal samples, added a pseudo-count of 1 to each mean and plotted the ratio between the means.

**Metabolomics analysis.** Osteosarcoma or melanoma cell lines were seeded at  $3 \times 10^6$  to  $5 \times 10^6$  cells per 10 cm plate and incubated with either 4 mM

L-glutamine, ( $\alpha$ - $^{15}\text{N}$ , 98%, Cambridge Isotope Laboratories) for 24 h. Subsequently, cells were washed with ice-cold saline, lysed with 50% methanol in water and quickly scraped followed by three freeze–thaw cycles in liquid nitrogen. The insoluble material was pelleted in a cooled centrifuge (4 °C) and the supernatant was collected for consequent GC–MS analysis. Samples were dried under air flow at 42 °C using a Techne Dry-Block Heater with sample concentrator (Bibby Scientific) and the dried samples were treated with 40  $\mu$ l of a methoxyamine hydrochloride solution (20 mg ml $^{-1}$  in pyridine) at 37 °C for 90 min while shaking followed by incubation with 70  $\mu$ l  $N,O$ -bis (trimethylsilyl) trifluoroacetamide (Sigma) at 37 °C for an additional 30 min.

**GC–MS.** GC–MS analysis used a gas chromatograph (7820AN, Agilent Technologies) interfaced with a mass spectrometer (5975 Agilent Technologies). An HP-5ms capillary column 30 m  $\times$  250  $\mu$ m  $\times$  0.25  $\mu$ m (19091S-433, Agilent Technologies) was used. Helium carrier gas was maintained at a constant flow rate of 1.0 ml min $^{-1}$ . The GC column temperature was programmed from 70 to 150 °C via a ramp of 4 °C min $^{-1}$ , 250–215 °C via a ramp of 9 °C min $^{-1}$ , 215–300 °C via a ramp of 25 °C min $^{-1}$  and maintained at 300 °C for an additional 5 min. The MS was by electron impact ionization and operated in full-scan mode from  $m/z$  = 30–500. The inlet and MS transfer line temperatures were maintained at 280 °C, and the ion source temperature was 250 °C. Sample injection (1  $\mu$ l) was in splitless mode. **Nucleic acid analysis.** *Materials.* Ammonium acetate (Fisher Scientific) and ammonium bicarbonate (Fluka) of LC–MS grade were used. Sodium salts of AMP, CMP, GMP, TMP and UMP were obtained from Sigma-Aldrich. Acetonitrile of LC grade was supplied from Merck. Water with resistivity 18.2 M $\Omega$  was obtained using Direct 3-Q UV system (Millipore).

*Extract preparation.* The obtained samples were concentrated in speedvac to eliminate methanol, and then lyophilized to dryness, re-suspended in 200  $\mu$ l of water and purified on polymeric weak anion columns (Strata-XL-AW 100  $\mu$ m (30 mg ml $^{-1}$ , Phenomenex)) as follows. Each column was conditioned by passing 1 ml of methanol, then 1 ml of formic acid/methanol/water (2/25/73) and equilibrated with 1 ml of water. The samples were loaded, and each column was washed with 1 ml of water and 1 ml of 50% methanol. The purified samples were eluted with 1 ml of ammonia/methanol/water (2/25/73) followed by 1 ml of ammonia/methanol/water (2/50/50) and then collected, concentrated in speedvac to remove methanol and lyophilized. Before LC–MC analysis, the obtained residues were re-dissolved in 100  $\mu$ l of water and centrifuged for 5 min at 21,000 g to remove insoluble material.

*LC–MS analysis.* The LC–MS/MS instrument consisted of an Acuity I-class UPLC system (Waters) and Xevo TQ-S triple quadrupole mass spectrometer (Waters) equipped with an electrospray ion source and operated in positive ion mode was used for analysis of nucleoside monophosphates. MassLynx and TargetLynx software (version 4.1, Waters) were applied for the acquisition and analysis of data. Chromatographic separation was done on a 100 mm  $\times$  2.1 mm internal diameter, 1.8- $\mu$ m UPLC HSS T3 column equipped with 50 mm  $\times$  2.1 mm internal diameter, 1.8- $\mu$ m UPLC HSS T3 pre-column (both Waters Acuity) with mobile phases A (10 mM ammonium acetate and 5 mM ammonium hydrocarbonate buffer, pH 7.0 adjusted with 10% acetic acid) and B (acetonitrile) at a flow rate of 0.3 ml min $^{-1}$  and column temperature 35 °C. A gradient was used as follows: for 0–6 min the column was held at 0% B, then 6–6.5 min a linear increase to 100% B, 6.5–7.0 min held at 100% B, 7.0–7.5 min back to 0% B and equilibration at 0% B for 2.5 min. Samples kept at 8 °C were automatically injected in a volume of 3  $\mu$ l.

For mass spectrometry, argon was used as the collision gas with a flow of 0.25 ml min $^{-1}$ . The capillary voltage was set to 2.90 kV, source temperature 150 °C, desolvation temperature 350 °C, desolvation gas flow 650 l min $^{-1}$ . Analytics were detected using multiple-reaction monitoring and applying the parameters listed in Supplementary Table 3.

**Hybridizations and imaging.** Single-molecule FISH (smFISH) was performed with probe libraries for Ass1 (74 probes, sequences described in Supplementary Methods) and Ki67 (96 probes<sup>20</sup>). Imaging was performed as previously described<sup>20</sup>. smFISH images were filtered with a Laplacian of Gaussian filter of size 15 pixels and standard deviation of 1.5 pixels. Each image is a maximum projection of ten stacks spaced 0.3  $\mu$ m apart in the  $z$ -direction. Each dot in these figures represents a cell and the quantification dots were counted on eight  $z$ -stacks spaced 0.3  $\mu$ m apart (total tissue volume 2.4  $\mu$ m).

**Proximity ligation assay.** The assay was performed as published<sup>21</sup> using Sigma Aldrich kit (DUO 92004-30-RXN). Antibodies used for detection were diluted in PBS: ASS1 (1:200, ab170952, abcam), citrin (1:100, H00010165-M01, clone 4F4, abnova) and anti-CAD (1:100, ab40800, abcam).

**Cell cultures.** All cell lines were authenticated. Melanoma cell lines LOX IMVI and MALME-3m and osteosarcoma cell lines MNNG/HOS and U2OS were purchased from American Type Culture Collection (ATCC) and cultured using standard

procedures in a 37°C humidified incubator with 5% CO<sub>2</sub> in RPMI (Invitrogen) supplemented with 10–20% heat-inactivated fetal bovine serum, 10% pen-strep and 2 mM glutamine. All cells are tested routinely for mycoplasma using a Mycoplasma EZ-PCR test kit (20–700–20, Biological Industries).

**Proliferation assays.** *MTT assay.* Cells were seeded in 12-well plates at  $4 \times 10^4$  to  $8 \times 10^4$  cells per well in a triplicate. After 6 h for adherence of the cells, 0.1 mg ml<sup>-1</sup> of MTT (3-(4,5-dimethylthiazol-2-yl)-2,5 diphenyltetrazolium bromide) (CAS 298-93-1, Calbiochem) in PBS was added to each cell type, starting at 0 h, in 24 h intervals. Deoxynucleotide Set (DNTP100-1KT, Sigma-Aldrich) was added to the cells' medium first after adherence and then daily at a final concentration of 10 µM. Cells were lysed with dimethylsulfoxide (DMSO). Absorbance was measured at 570 nm.

**Crystal violet staining.** Cells were seeded in 12-well plates at 40,000–100,000 cells per well in a triplicate. Time 0 was calculated as the time the cells became adherent, which was after about 6 h from plating. For each time point, cells were washed with PBS X1 and fixed in 4% PFA (in PBS). Cells were then stained with 0.1% Crystal Violet (C0775, Sigma-Aldrich) for 20 min (1 ml per well) and washed with water. Cells were then incubated with 10% acetic acid for 20 min with shaking. Extract was then diluted 1:4 in water and absorbance was measured at 590 nm every 24 h.

**Protein and RNA analysis.** *Western blotting.* Cells were lysed in RIPA (Sigma-Aldrich) and 0.5% protease inhibitor cocktail (Calbiochem). After centrifugation, the supernatant was collected and protein content was evaluated by the Bradford assay. One hundred micrograms from each sample under reducing conditions were loaded into each lane and separated by electrophoresis on a 10% SDS polyacrylamide gel. After electrophoresis, proteins were transferred to Immobilon transfer membranes (Tamar). Non-specific binding was blocked by incubation with TBST (10 mM Tris-HCl (pH 8.0), 150 mM NaCl, 0.1% Tween 20) containing 3% albumin from bovine serum for 1 h at 25°C. Membranes were subsequently incubated with antibodies against ASS1 (1:500, sc-99178, Santa Cruz Biotechnology)<sup>22</sup>, p97 (1:10,000, PA5-22257, Thermo Scientific), GAPDH (1:1,000, 14C10, 2118, Cell Signaling)<sup>23</sup>, CAD (1:1,000, ab40800, abcam)<sup>24</sup>, phospho-CAD (Ser1859) (1:1,000, 12662, Cell Signaling)<sup>15</sup>, p70 S6 Kinase (1:1,000, 9202, Cell Signaling) and phospho-p70 S6 Kinase (Ser371) (1:1,000, 9208, Cell Signaling)<sup>25</sup>. Antibody was detected using peroxidase-conjugated AffiniPure goat anti-rabbit IgG or goat anti-mouse IgG (Jackson ImmunoResearch) and enhanced chemiluminescence western blotting detection reagents (EZ-Gel, Biological Industries).

Gels were quantified by Gel Doc XR+ (BioRad) and analysed by ImageLab 4.1 software (BioRad). The band area was calculated by the intensity of the band. The obtained value was then divided by the value obtained from the loading control.

**RNA extraction and complementary DNA (cDNA) synthesis.** RNA was extracted from cells by using PerfectPure RNA Cultured Cell Kit (5'-PRIME). cDNA was synthesized from 1 µg RNA by using qScript cDNA Synthesis Kit (Quanta).

**Quantitative PCR.** Detection of ASS1 on cDNAs (see above) was performed using SYBR green PCR master mix (Tamar) and the required primers. Primer sequences were as follows. Human ASS1: forward, 5'-TTATAACCTGGGATGGGCACC-3'; reverse, 5'-TGGACATAGCGTCTGGGATTG-3'. Human HPRT: forward, 5'-ATTGACACTGGCAAAACAATGC-3'; reverse, 5'-TCCAACACTTCG TGGGGTCC-3'. Analysis used StepOne real-time PCR technology (Applied Biosystems).

**Transient transfection.** Cells were seeded in 12-well plates at 30,000 cells per well, or in 10 cm plates at  $10^6$  cells per plate, in triplicate. The following day, cells were transfected with either 20 pmol or 600 pmol siRNA siGenome SMARTpool targeted to Citrin mRNA (M-007472-01, Thermo Scientific), respectively. Transfection was performed with Lipofectamine 2000 Reagent (11668-019, Invitrogen) in the presence of Opti-MEM1 Reduced Serum Medium (31985-062, Invitrogen). Four hours after transfection, medium was replaced and experiments were performed starting 24 h after transfection.

**Infection.** Over-expression. Cells were infected with pLenti3.3/TR and with pLenti6.3/TO/V5-DEST-based lentiviral vector with or without the human ASS1 transcript. Transduced cells were selected with 1 mg ml<sup>-1</sup> Geneticin and with 7.5 µg ml<sup>-1</sup> Blastidin for each plasmid, respectively. When induction of expression was needed, cells were added with 10 µg ml<sup>-1</sup> tetracycline/doxycycline.

**Short hairpin RNA.** Cells were infected with pLKO-based lentiviral vector with or without the human ASS1 short hairpin RNA (shRNA) encoding one or two separate sequences combined (RHS4533-EG445, GE Healthcare, Dharmacon). Transduced cells were selected with 2 µg ml<sup>-1</sup> puromycin.

**Arginine deprivation combined with drug treatments.** U2OS human osteosarcoma cell line was seeded in 6-well plates at 80,000 cells per well. The following

day, cells were treated with either 100 nM rapamycin (R0395, Sigma-Aldrich) or with 10 µM 5FU (F6627, Sigma-Aldrich) in regular medium, with 10% dialysed FCS-arginine-free-RPMI (06-1104-34-1A, Biological Industries) or with both arginine-depleted medium and one of these drugs. Rapamycin and 5FU were renewed into the medium every day, whereas fresh arginine-free medium was supplemented twice a week.

**Animal studies.** According to the approved IACUC protocol 17270415-2, tumours did not exceed the limits of more than 10% of the animal weight and were not longer than 1.5 cm in length in any dimension (Supplementary Fig. 2). Ten million MALME-3m melanoma cells suspended in 500 µl PBS with 5% Matrigel (4132053 Corning) were injected subcutaneously to 8- to 12-week-old male SCID mice that were purchased from Harlan. There were 22 SCID mice, from which 5 or 6 were used for each cell line in each of the three experiments performed. No randomization was used. Mice were monitored for survival and tumour burden twice a week by a veterinarian investigator who was blinded to the expected outcome. Tumours were measured using a calliper. After euthanization, tumours were removed and incubated in medium containing [<sup>15</sup>N]glutamine for 6 h followed by GC-MS analysis. Tumour size was calculated as published<sup>26</sup>.

**Building cell models.** We used genome-scale metabolic models of NCI-60 cancer cell lines. The reconstruction method (on the basis of methods termed PRIME<sup>10</sup> requires several key inputs: (1) the generic human model<sup>7</sup>; (2) gene expression data for each cell line<sup>19</sup>; and (3) growth rate measurements (available at the NCI website: [https://dtp.cancer.gov/discovery\\_development/nci-60/cell\\_list.htm](https://dtp.cancer.gov/discovery_development/nci-60/cell_list.htm)). The algorithm then reconstructs a specific metabolic model for each sample by modifying the upper bounds of reactions in accordance with the expression of the individual gene microarray values.

Specifically, the model reconstruction process is as follows. (1) Decompose reversible reactions into unidirectional forward and backward reactions. (2) Evaluate the correlation between the expression of each reaction in the network and the measured growth rate. The expression of a reaction is defined as the mean over the expression of the enzymes catalysing it. (3) Modify upper bounds on reactions demonstrating significant correlation to the growth rate (after correcting for multiple hypothesis using false discovery rate) in a manner that is linearly related to expression value.

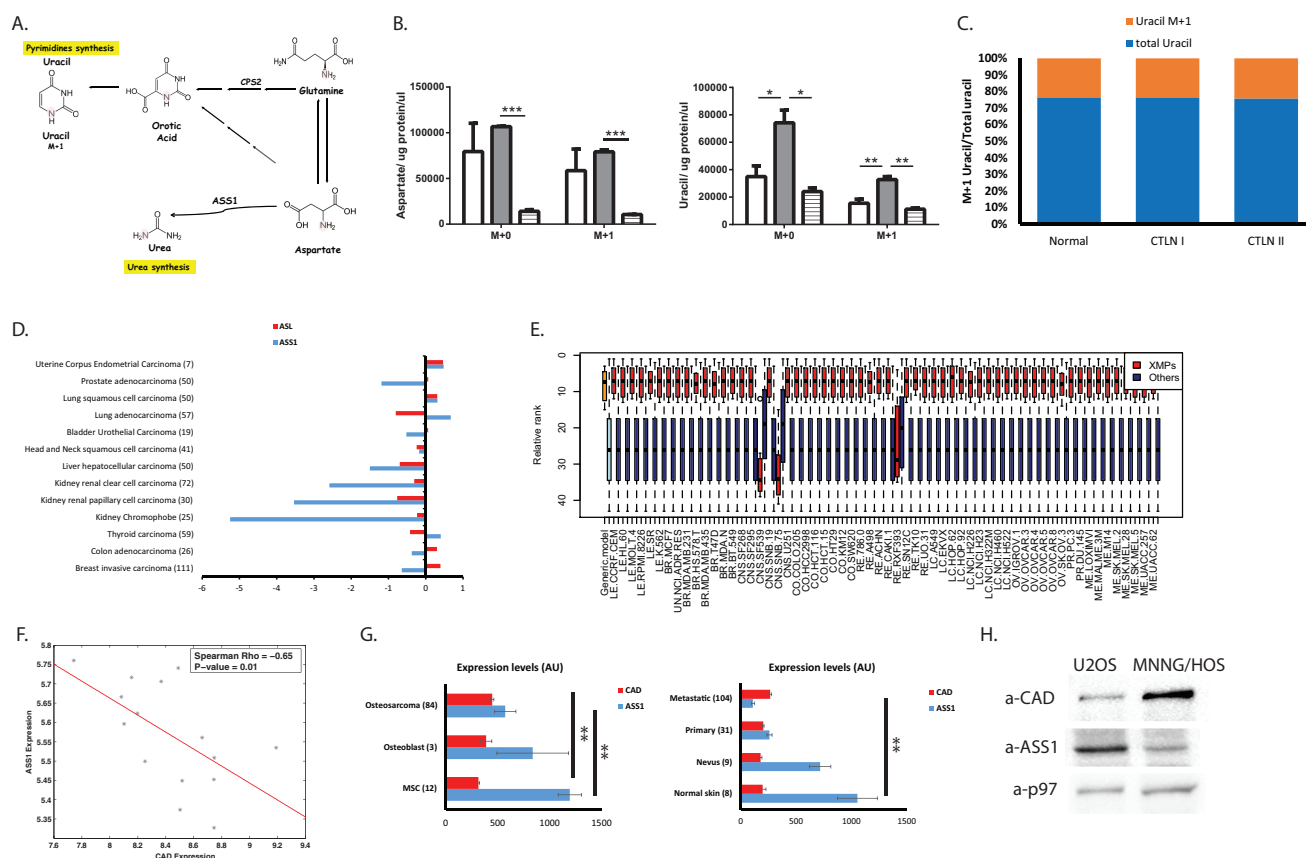
**Ass1 probes.** Gcgatcagggttataagc; gcggagcagggtgagagag; gccagggcggcagtg gaga; ggccagatgaaccactcagt; agtcctgtgtagctgcttc; gattataggtacaggtccct; cttgtctg gacatctgtct; ctgtaggccagacaaccacaga; gcaggaggtgtccaggccac; gttccttcagccacagagg; taggcgatgacatcatagcc; cttctggcaatgttgccca; tcttgcttctcaaatgt; gccccaagcttcag cgctt; atctcaatgaacacctttt; cttccacaatctctgtct; tggacagcaggccagatgaa; gttcctgta gaggctcagtg; gaggaggtgccaggagatag; cgagctatgcaaggcctgg; ctgggcaatctccacct gtc; acacatactggcccctca; cctttccctggcgccgctg; ctcaaggcgacgtggtcat; gtgcaggt gaatagcaggtg; ggagcgatgaccttaactg; tgttaaacctcaggcatcct; atcatttcggcccttggaac; gttccttcgatactccatc; gtgacagggatggggatgcc; catactccaggggctcttg; tgatgtgcatgag gtttca; tccaggatcccagcctcata; agtgcttgattcttggtg; gaggttttgtgtagagacc; ttgg gtgcttgccaggggt; tatctcaaggacatctggg; cagggacccctttttgaa; tcttgatgttggtacact; ggatgtggtgctgggtgtg; tcaggatcatgaagagttcc; ccgtgcttgcccgcaactc; cagcatgtcaatg gcacca; tcatctcaatgaagcgggt; gttcctgtagatacctcgga; gtaaggatgggtccctgctg; cctc tatgtctaatgagcgg; acttccgatccatctgtgaa; caggccctgctgtatt; cgagctctgcgaatttgagg; ctgtgccagaaacctgtgta; gcaacaaattcacattcag; cctgggactcttgatacag; tgcaccttccctc tacccg; ttggcccttgaaagacagaca; actccgaccaggagatgac; tcatgtgagtgaaagtgg; tgcaggt tcatgctcaccag; gtcgatgggctcatagtcg; tgatattgatgaagccagtg; tactccttcagcctgagcga; gacctgtctgaaggcgat; ttgtcagggtctatttgga; gagggtggaggcccgctcct; gctgaagcctggga gactg; caaatttatcacaacaa; ggtggagaacaagctacaat; gacacagcagcccgatcag; aggcgtg gggggggcgggg; gctataggggaccagggaac; ccttgatgaccactttgt; agctcccgccaccctccct; attgtcattttatgcttct; aagactaatgtaacttctt.

**Statistics.** All statistical analyses were performed using Tukey's honest significant difference test or independent-samples Student's *t*-test of multiple or two groups, respectively. Log-transformed data were used where differences in variance were significant and variances were correlated with means. The sample size was chosen in advance on the basis of common practice of the described experiment and is mentioned for each experiment. No statistical methods were used to predetermine sample size. Each experiment was conducted with biological and technical replicates and repeated at least three times unless specified otherwise. On the basis of pre-established criteria, individual outlier data points that were more than 2 standard deviations away from the mean were excluded from the data analysis. Statistical tests were done using Statsoft's STATISTICA, version 10. All error bars are standard errors.  $P < 0.05$  was considered significant in all analyses ( $^*P < 0.05$ ,  $^{**}P < 0.005$ ,  $^{***}P < 0.0005$ ).

**Kaplan–Meier.** For each cancer type, the Kaplan–Meier plot indicates the survival rates of the four different groups of patients as labelled. We analysed the cancer types for which there were sufficient survival data.



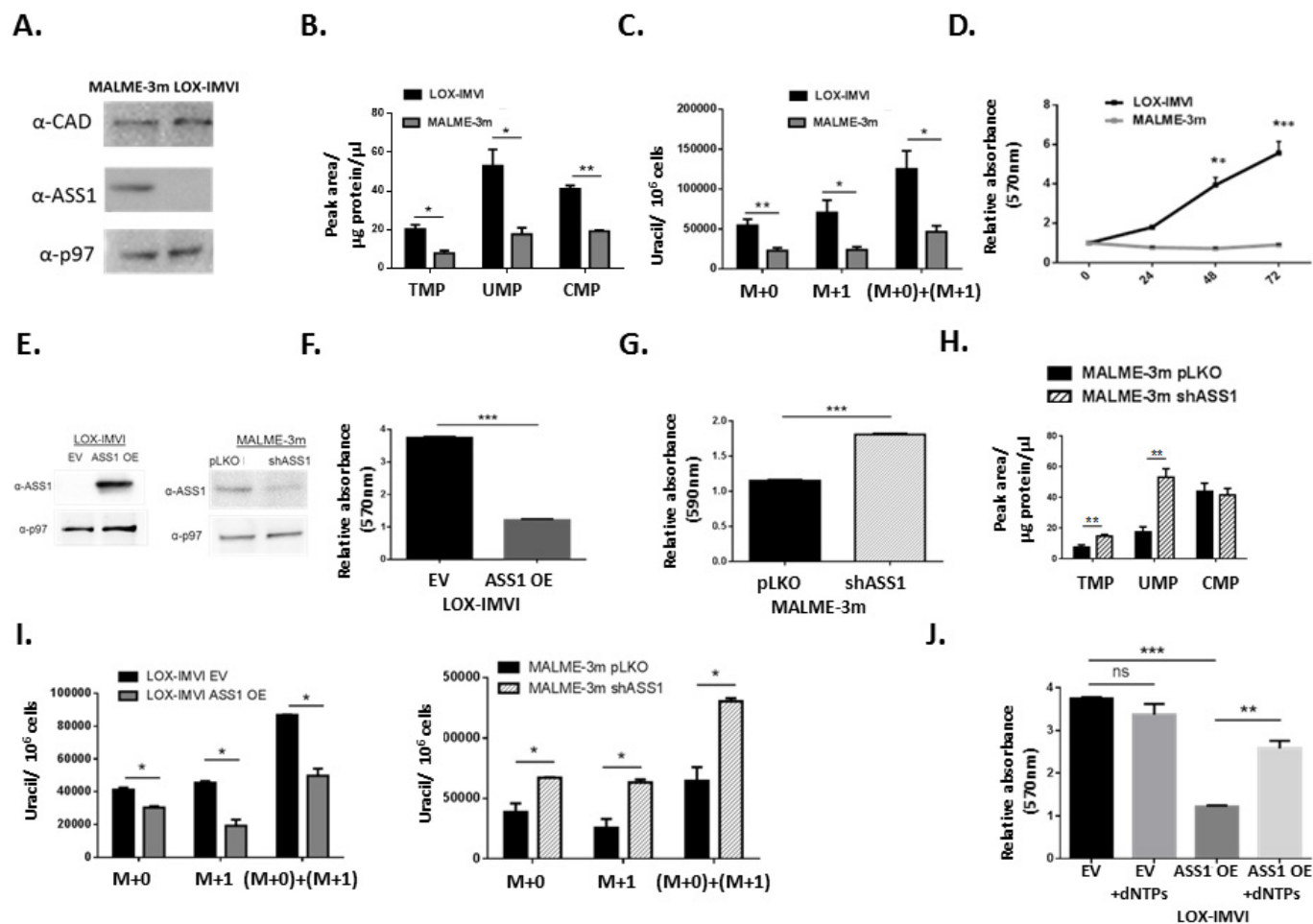
16. Folger, O. *et al.* Predicting selective drug targets in cancer through metabolic networks. *Mol. Syst. Biol.* **7**, 501 (2011).
17. Varma, A. P. & Palsson, B. O. Metabolic flux balancing: basic concepts, scientific and practical use. *Nature Biotechnol.* **12**, 994–998 (1994).
18. Bordel, S., Agren, R. & Nielsen, J. Sampling the solution space in genome-scale metabolic networks reveals transcriptional regulation in key enzymes. *PLOS Comput. Biol.* **6**, e1000859 (2010).
19. Lee, J. K. *et al.* A strategy for predicting the chemosensitivity of human cancers and its application to drug discovery. *Proc. Natl Acad. Sci. USA* **104**, 13086–13091 (2007).
20. Itzkovitz, S. *et al.* Single-molecule transcript counting of stem-cell markers in the mouse intestine. *Nature Cell Biol.* **14**, 106–114 (2012).
21. Gu, G. J. *et al.* Protein tag-mediated conjugation of oligonucleotides to recombinant affinity binders for proximity ligation. *New Biotechnol.* **30**, 144–152 (2013).
22. Hao, G., Xie, L. & Gross, S. S. Argininosuccinate synthetase is reversibly inactivated by S-nitrosylation *in vitro* and *in vivo*. *J. Biol. Chem.* **279**, 36192–36200 (2004).
23. Bjorklund, N. L., Sadagoparamanujam, V. M. & Taglialetela, G. Selective, quantitative measurement of releasable synaptic zinc in human autopsy hippocampal brain tissue from Alzheimer's disease patients. *J. Neurosci. Methods* **203**, 146–151 (2012).
24. Robitaille, A. M. *et al.* Quantitative phosphoproteomics reveal mTORC1 activates de novo pyrimidine synthesis. *Science* **339**, 1320–1323 (2013).
25. Zhang, Y. *et al.* Signal transduction pathways involved in phosphorylation and activation of p70S6K following exposure to UVA irradiation. *J. Biol. Chem.* **276**, 20913–20923 (2001).
26. Pinthus, J. H. *et al.* WISH-PC2: a unique xenograft model of human prostatic small cell carcinoma. *Cancer Res.* **60**, 6563–6567 (2000).



### Extended Data Figure 1 | ASS1 deficiency correlates with aspartate utilization by CAD in cancerous and non-cancerous cells.

**a.** Schematic flux tracing of the  $\alpha$ -labelled nitrogen of glutamine ( $[^{15}\text{N}]$   $\alpha$ -glutamine) to nucleic acid synthesis via aspartate. **b.** The ratio between M+1-labelled and total uracil in fibroblasts is similar between patients with citrullinaemia patients and control subjects ( $n \geq 3$ ). Error bars are standard error. **c.** Labelled levels of M+1 aspartate (left) and M+1 uracil (right) synthesized from  $[^{15}\text{N}]$   $\alpha$ -glutamine are higher in fibroblasts from CTLN I than in fibroblasts from controls and patients with CTLN II ( $n \geq 3$ ). **d.** TCGA analysis of tumour–normal paired tissues for gene expression comparison shows the expression levels of *ASL* and *ASS1* in different cancers. **e.** Plot generated from the modelling data for the production capacity of metabolites after *ASS1* inactivation in each of

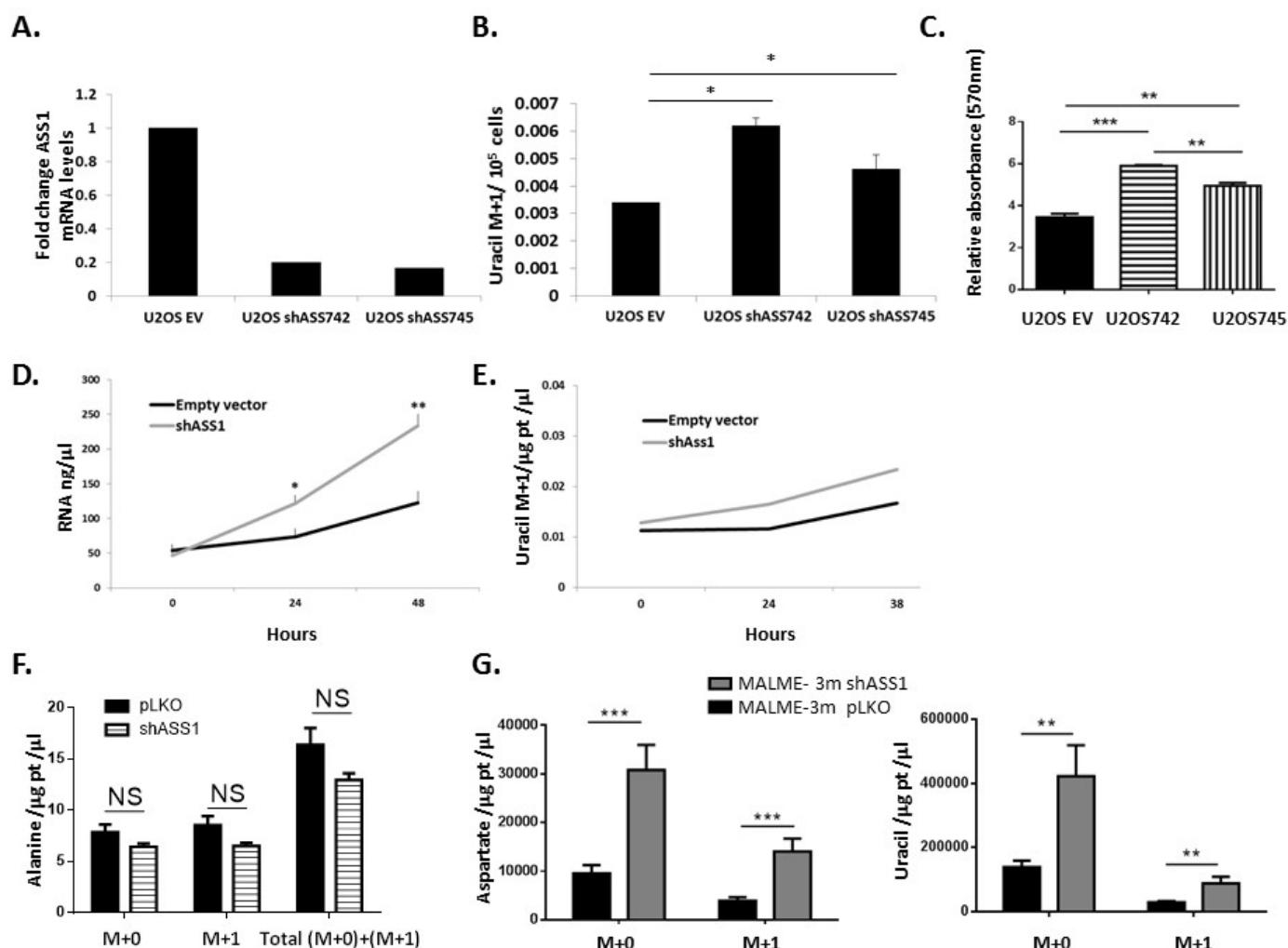
the NCI-60 cell lines as well as in the generic model. The reddish bars represent the ranking of nucleic acids while the blueish bars represent the ranking of all other metabolites. **f.** Correlation analysis of NCI-60 cell lines shows a significant inverse correlation between *ASS1* and *CAD* expression levels. **g.** Osteosarcoma (upper) and melanoma (lower) microarray data was obtained from the NCBI EO database (accession numbers GSE33383 and GSE46517, respectively). Raw expression levels were plotted and significance was computed using a *t*-test on  $\log_2$ -transformed expression levels. The number of patients for each subtype is shown in parenthesis on the left. **h.** Western blot for *CAD* and *ASS1* shows higher expression levels of *CAD* in the MNNG/HOS human osteosarcoma cell line, which has a low expression level of *ASS1* compared with U2OS, which has higher expression levels of *ASS1*; p97 is shown as loading control.



**Extended Data Figure 2 | ASS1 inactivation in melanoma correlates with increased proliferation.** **a**, An immunoblot showing different expression levels of ASS1 and CAD in two different cancer cell lines of melanoma. **b**, Melanoma cells with ASS1 downregulation have a significant increase in pyrimidine levels as measured by LC-MS ( $n \geq 3$ ). **c**, Melanoma cells with ASS1 downregulation have a significant increase in total uracil ( $n = 4$ ). **d**, Melanoma cells with ASS1 downregulation have a significant increase in proliferation as measured by MTT assay ( $n = 2$ ). **e**, Immunoblots of melanoma cells for ASS1 levels after transduction with either ASS1 over expression construct or with *shASS1*.

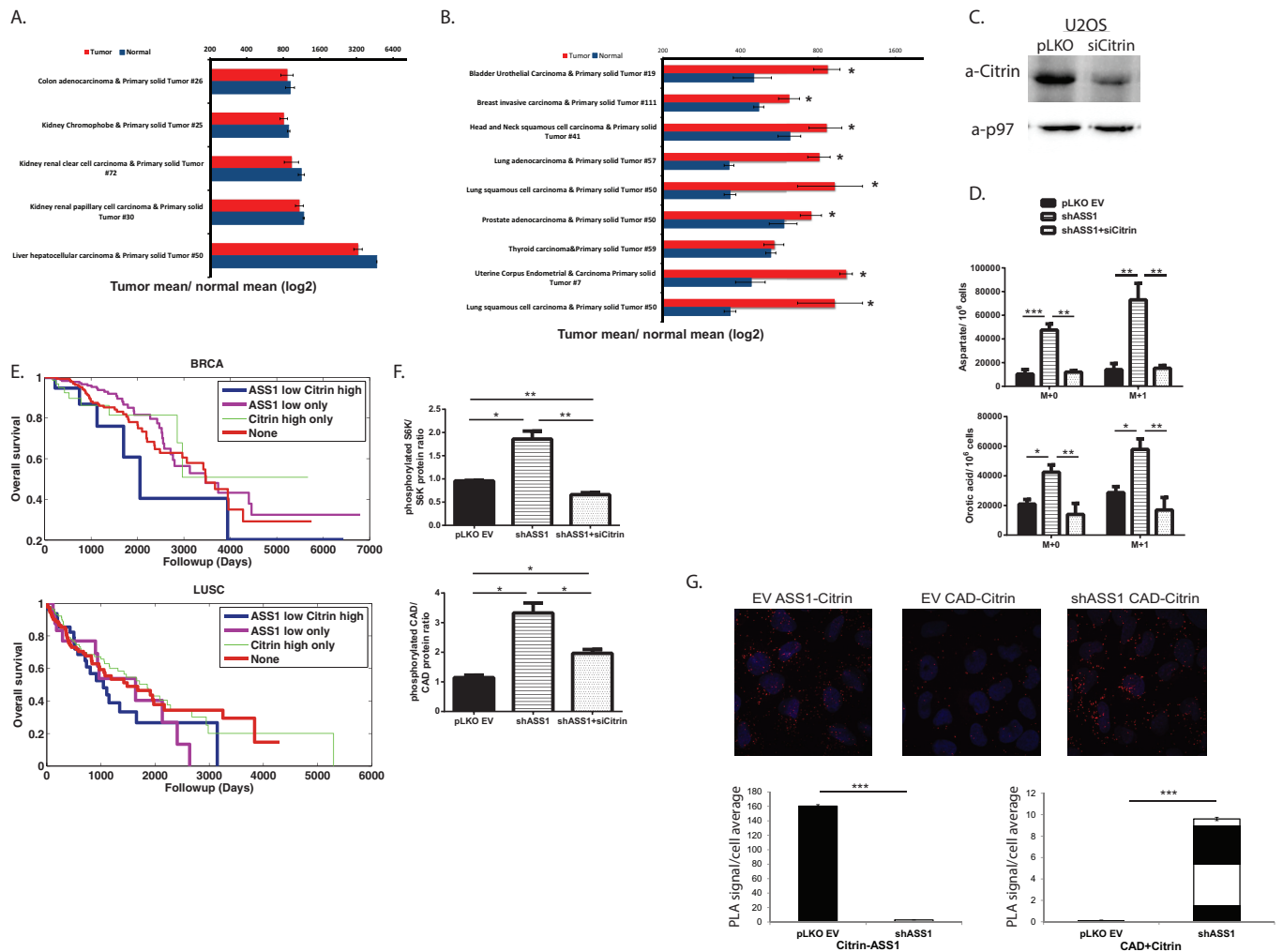
**f**, Proliferation assays showing a significant decrease in proliferation after ASS1 overexpression in melanoma using MTT ( $n = 3$ ). **g**, Crystal violet quantification for melanoma cells after transduction with *shASS1* demonstrating increase in proliferation ( $n = 3$ ). **h**, LC-MS measurements of pyrimidine levels showing a significant increase after the use of *shASS1* in melanoma cells ( $n \geq 3$ ). **i**, Left: total uracil levels are decreased significantly in melanoma cells with ASS1 overexpression and increased in melanoma cells with *shASS1* (right) ( $n \geq 2$ ). **j**, Significant increase in proliferation of melanoma cells by dNTPs after ASS1 overexpression ( $n = 3$ ). All error bars are standard errors.





**Extended Data Figure 3 | Downregulation of ASS1 levels increases pyrimidine synthesis.** **a**, Osteosarcoma cells were transduced with two different *shASS1* vectors: *shASS742* and *shASS745*. Both clones decreased ASS1 levels efficiently to approximately 20% expression (left), resulting in a significant increase in uracil M + 1 levels (**b**) and in proliferation ( $n \geq 3$ ) (**c**). **d**, RNA levels measured in U2OS at 24 h intervals show increased levels of RNA in U2OS infected with *shASS1* compared with the empty

vector. **e**, Uracil M + 1 levels increase more in U2OS infected with *shASS1* compared with the empty vector during 38 h of measurements. **f**, The levels of total and labelled M + 1 alanine synthesis from [<sup>15</sup>N]α-glutamine do not change significantly after ASS1 downregulation ( $n = 3$ ). **g**, Tumours with *shASS1* had higher levels of M + 1 aspartate (left) and M + 1 uracil (right) synthesized from [<sup>15</sup>N]α-glutamine, compared with tumours expressing the empty vector ( $n = 15$ ). All error bars are standard errors.



**Extended Data Figure 4 | Cancers with ASS1 downregulation are addicted to aspartate.** **a, b,** Analysis of the TCGA database of matched tumour–normal pairs showing no significant difference in the expression level of *citrin* in tissues with a high baseline expression of *citrin* (**a**) and significant elevation in tumours in which the normal tissue has low basal expression of *citrin* (**b**) ( $*P < 0.001$ ). **c,** Immunoblot showing the expression level of citrin in osteosarcoma cells after *si-citrin*. **d,** Labelled and unlabelled aspartate (top) and uracil (bottom) are elevated significantly in cancers with ASS1 downregulation and are comparable to control in cells with both ASS1 and citrin downregulation ( $n \geq 3$ ). Error bars are standard errors. **e,** Kaplan–Meier survival analysis for two different cancer types (BRCA, breast cancer (top); LUSC, lung squamous cell carcinoma (bottom)), showing poor survival trend for cancers with low ASS1 and high citrin. For each cancer type, the Kaplan–Meier plot

indicates the survival rates of four groups of patients: (1) ASS1 low expression and citrin high expression; (2) ASS1 low expression; (3) citrin high expression; (4) none of these. We analysed the cancer types for which there were sufficient survival data. **f,** Quantification graph of a western blot showing decreased CAD and S6K phosphorylation after treatment of U2OS with *si-citrin*. Error bars are standard errors. **g,** Proximity ligation assay showing increased proximity between CAD and citrin after ASS1 knockdown in U2OS cells (top, red dots). The left and middle pictures show the proximity between ASS1 and CAD to citrin in U2OS infected with empty vector, whereas the right picture shows the proximity between CAD and citrin after infection of U2OS with shASS1. Bottom: quantification of proximity ligation assays performed on U2OS infected with either empty vector (EV) or with *shASS1* using antibodies for citrin, ASS1 and CAD. The pictures were quantified using ImageJ.

Extended Data Table 1 | ASS1 inactivation is predicted to increase aspartate flux for nucleic acid synthesis

Aspartate			
Metabolic Pathway	Catalyzing Enzymes	Inactive ASS1/Active ASS1	P-value
Pyrimidine Biosynthesis	(790.1), CAD	↑	8.47E-198
IMP Biosynthesis	(10606.1) PAICS	↑	<1e-300
Nucleotides	(159.1) Adenylosuccinate synthase	↑	1.55E-265

Predicted fold-change in flux rates through pathways associated with aspartate and glutamine, when comparing ASS1 inactivation with activation state. The most significant change is predicted to effect the pyrimidine biosynthesis pathway followed by purine synthesis pathway (two-sided Wilcoxon rank-sum  $P$  value  $< 8.4 \times 10^{-198}$ , Methods).



Extended Data Table 2 | Kaplan–Meier log-rank data analysis shows significant worsening in the survival of patients with low ASS1 and high citrin expression levels in bladder cancer and lung adenocarcinoma

BLCA		Number of patients	ASS1 low only	Citrin high only	None
	ASS1 low Citrin high	26	0.062029244	0.41802416	0.002920577
	ASS1 low only	437	0	0.939396939	0.133262071
	Citrin high only	83	0	0	0.397997761
	None	415			
BRCA		Number of patients	ASS1 low only	Citrin high only	None
	ASS1 low Citrin high	25	0.091470414	0.348538412	0.215689269
	ASS1 low only	76	0	0.513641931	0.247876216
	Citrin high only	14	0	0	0.784222648
	None	90			
LUAD		Number of patients	ASS1 low only	Citrin high only	None
	ASS1 low Citrin high	34	0.086250611	0.000771207	0.005978456
	ASS1 low only	16	0	0.989469401	0.778260123
	Citrin high only	181	0	0	0.440612674
	None	189			
LUSC		Number of patients	ASS1 low only	Citrin high only	None
	ASS1 low Citrin high	56	0.889254024	0.131597923	0.271250284
	ASS1 low only	35	0	0.234416519	0.540649678
	Citrin high only	173	0	0	0.710240332
	None	127			

The table shows the number of patients for whom data were available in each group, as well as the pairwise *P* value of comparison between the corresponding groups: BLCA, BRCA, LUAD, LUSC.

# DNA-dependent formation of transcription factor pairs alters their binding specificity

Arttu Jolma<sup>1</sup>, Yimeng Yin<sup>1</sup>, Kazuhiro R. Nitta<sup>1</sup>, Kashyap Dave<sup>1</sup>, Alexander Popov<sup>2</sup>, Minna Taipale<sup>1</sup>, Martin Enge<sup>1</sup>, Teemu Kivioja<sup>3</sup>, Ekaterina Morgunova<sup>1</sup> & Jussi Taipale<sup>1,3</sup>

Gene expression is regulated by transcription factors (TFs), proteins that recognize short DNA sequence motifs<sup>1–3</sup>. Such sequences are very common in the human genome, and an important determinant of the specificity of gene expression is the cooperative binding of multiple TFs to closely located motifs<sup>4–6</sup>. However, interactions between DNA-bound TFs have not been systematically characterized. To identify TF pairs that bind cooperatively to DNA, and to characterize their spacing and orientation preferences, we have performed consecutive affinity-purification systematic evolution of ligands by exponential enrichment (CAP-SELEX) analysis of 9,400 TF–TF–DNA interactions. This analysis revealed 315 TF–TF interactions recognizing 618 heterodimeric motifs, most of which have not been previously described. The observed cooperativity occurred promiscuously between TFs from diverse structural families. Structural analysis of the TF pairs, including a novel crystal structure of MEIS1 and DLX3 bound to their identified recognition site, revealed that the interactions between the TFs were predominantly mediated by DNA. Most TF pair sites identified involved a large overlap between individual TF recognition motifs, and resulted in recognition of composite sites that were markedly different from the individual TF's motifs. Together, our results indicate that the DNA molecule commonly plays an active role in cooperative interactions that define the gene regulatory lexicon.

The set of rules by which a DNA sequence can be converted into knowledge of spatial and temporal expression patterns of a protein has been difficult to decipher<sup>1–3</sup>. This is in part because, in mammals, more than 1,000 TFs recognizing over 200 different short DNA motifs participate in interpreting gene regulatory information<sup>7–10</sup>. In addition, TFs also interact with each other, and many TFs bind DNA as homo- or heterodimers. A pair of TFs can bind to multiple different DNA motifs, as the recognition sites of individual TFs can occur in different orientations and/or spacings relative to each other<sup>11–13</sup>. Most known heterodimeric interactions occur between two TFs of the same structural family, but several cases where TFs of different structural classes bind cooperatively have also been identified<sup>4,6</sup>.

To chart the prevalence of co-operative interactions between TFs in the presence of DNA, we developed a novel method, CAP-SELEX, in which specific DNA sequences that interact with two different TFs at the same time are selected from a library of random sequences. CAP-SELEX only detects a specific type of interaction involving three macromolecules, where the two tested proteins both bind to the same DNA in a sequence-specific manner (Fig. 1a). Compared to existing methods such as SELEX-seq<sup>14</sup> and universal protein binding microarrays<sup>7,15</sup>, CAP-SELEX respectively allows higher throughput and interrogation of larger sequence space. A total of 100 streptavidin-binding peptide (SBP) tagged TF1 proteins were used in the assay together with 94 (3 × Flag-tagged) TF2 proteins to characterize 9,400 potential interactions (Fig. 1a and Supplementary Table 1). TFs were selected to cover a wide variety of structural classes and individual binding specificities.

To allow bacterial expression, unstructured regions and amino- and carboxy-terminal sequences that do not correspond to known protein domains were removed from the constructs.

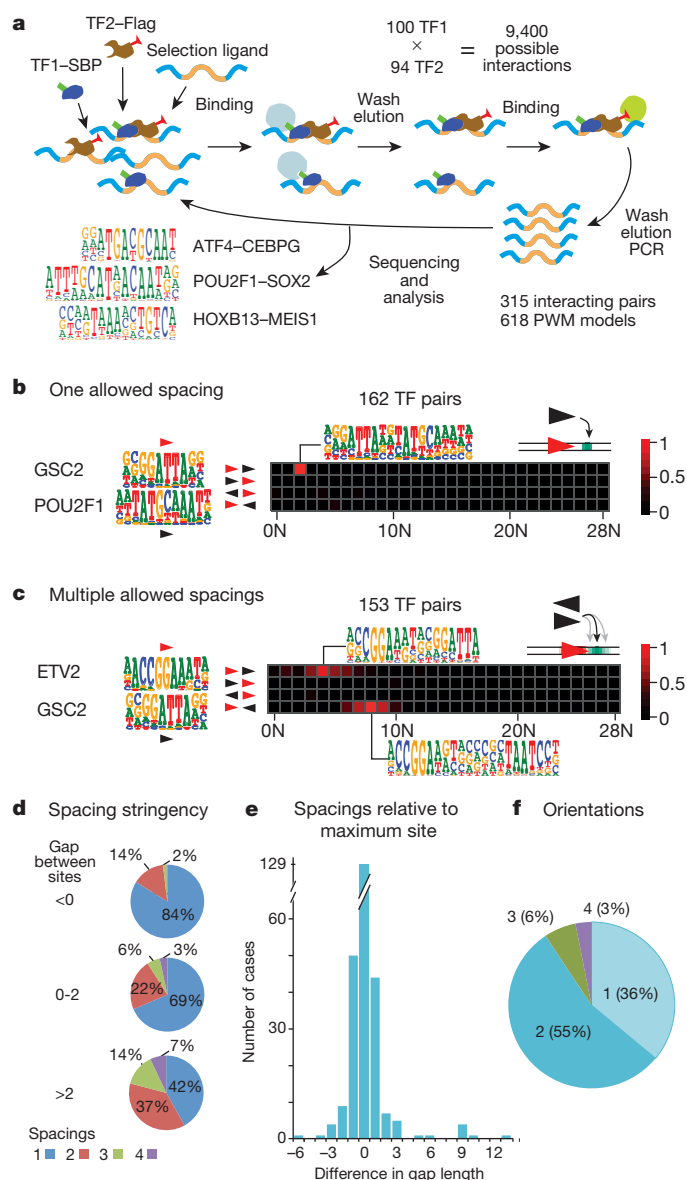
Co-operative signals were detected from CAP-SELEX enriched sequences (Extended Data Fig. 1) either as novel composite sites that combine partial specificities of the two TFs (using Autoseed<sup>16</sup>), or as orientation and spacing preferences between the individual TF's motifs (Supplementary Table 2). This analysis revealed that 55% (55/100) of TF1 and 70% (66/94) of TF2 proteins were 'active' in the CAP-SELEX assay, as indicated by identification of the expected pair of monomeric sites in at least five experiments, or a heterodimeric site in at least one experiment (Extended Data Fig. 2 and Supplementary Table 1). To test reproducibility of the assay, 10 TF1 proteins were run again against all TF2 proteins. In most cases, the recovered motifs were very similar. In addition, we validated 10 TF–TF pairs using purified full-length proteins (Extended Data Fig. 3).

Of the 3,630 tested interactions between active TF1–TF2 pairs, 315 (8.7%) displayed cooperative binding. This result is likely an underestimate, as enrichment of both expected motifs was not observed in many cases (83% of all tested pairs, not shown). The interactions were not limited to those between related TFs, and also occurred commonly between TFs from different structural families. Only 5% of all active TF1 and TF2 pairs appeared to bind to DNA independently of each other, as indicated by the presence of both expected motifs without strong orientation and spacing preferences (Extended Data Fig. 2).

Of the interacting TF pairs, 162 had only one preferred site, whereas 153 pairs displayed more than one spacing and/or orientation (Fig. 1b, c). Analysis of pairs of motifs that occurred in the same orientation revealed that the stringency of their spacing was dependent on the motif-to-motif distance. Most TF pairs whose motifs overlapped preferred just one (negative) spacing between the motifs. In contrast, if the most enriched motif pair had a gap, two or more spacings were more commonly observed. Most longer-range interactions where the gap between the motifs was 3 base pairs (bp) or more displayed a relatively wide,  $\pm 2$  bp, spacing preference, similar to that reported previously<sup>17</sup> (Fig. 1d and Extended Data Fig. 4). Many TF pairs displayed both kinds of interactions, with one orientation preferring stringent short-range interactions, and the other orientation(s) preferring the more relaxed long-range interactions (not shown).

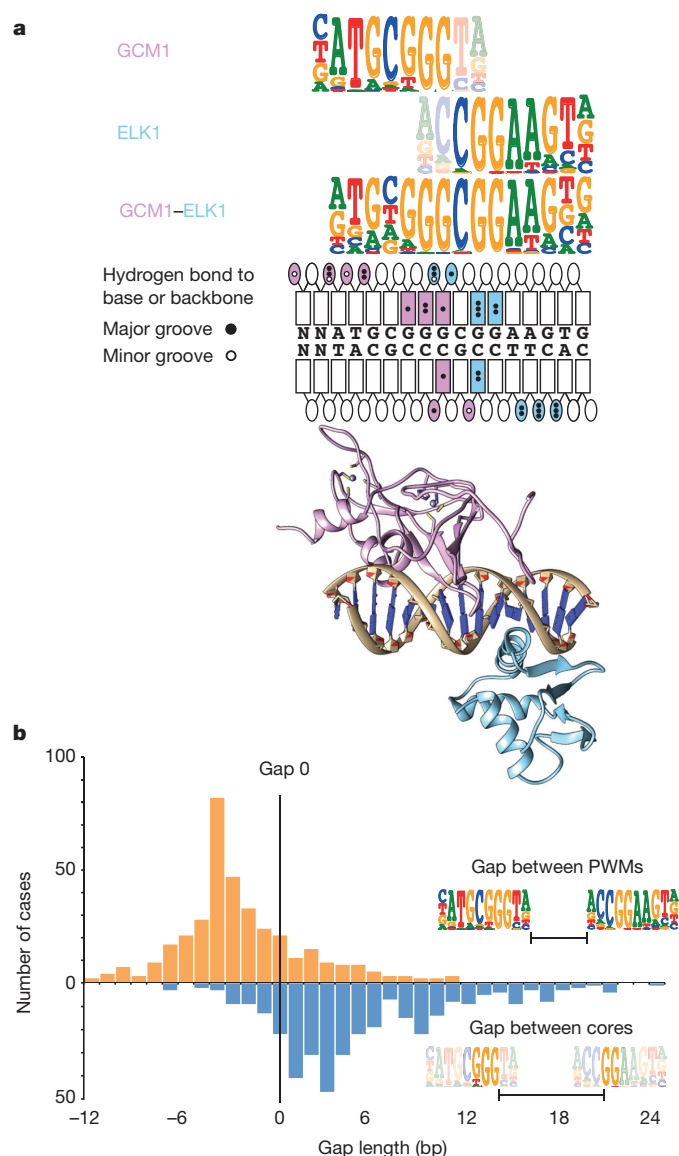
In cases where two or more motif spacings were allowed in the same orientation, the differences between the observed motif-to-motif distances were in general very small (Fig. 1e, >73% were only 1 bp), indicating that the mechanism of the TF–TF cooperativity is very sensitive to the relative distance and/or angle between the bound TFs. The promiscuous nature of the cooperativity was highlighted by the fact that most cases where more than one preferred mode of binding was observed involved multiple motif orientations (Fig. 1f). Two orientations was most common, whereas fewer cases of three or four orientations were observed, in part because pairs with one or two TFs with

<sup>1</sup>Department of Biosciences and Nutrition, Karolinska Institutet, SE 141 83, Sweden. <sup>2</sup>European Synchrotron Radiation Facility, 38043 Grenoble, France. <sup>3</sup>Genome-Scale Biology Program, University of Helsinki, P.O. Box 63, FI-00014, Finland.



**Figure 1 | CAP-SELEX reveals DNA-mediated TF-TF interactions.** **a**, Schematic description of CAP-SELEX. A TF1-TF2-DNA complex is formed (top left) and subjected to two consecutive affinity purifications, followed by amplification of DNA and sequencing. The entire process is repeated three times, and the cooperative complexes are then detected from the sequences. CAP-SELEX derived PWM models for the indicated previously known<sup>26-28</sup> TF complexes are also shown. **b**, An example of a TF-TF pair preferring a single spacing and orientation. Heatmap shows counts (divided by max) of representative 6-mers for the TFs. **c**, A TF pair preferring two different orientations, with relatively flexible spacing in both orientations. Logos for the strongest cases are also shown. **d**, High stringency of closely packed TF-TF sites. The pie charts show the number of allowed spacings in a single orientation, binned according to the motif-to-motif distance (gap) of the strongest-bound site (maximum). **e**, TFs that can bind to sites with multiple spacings prefer very closely spaced sites. Histogram shows difference in gap length between the most strongly enriched motif spacing (normalized to 0) and other identified motif spacings. **f**, Most TF-TF pairs with multiple cooperatively bound sites allow more than one orientation. Pie chart shows frequency of TF-TF pairs binned according to the number of allowed orientations. Only pairs with multiple preferred motifs are included.

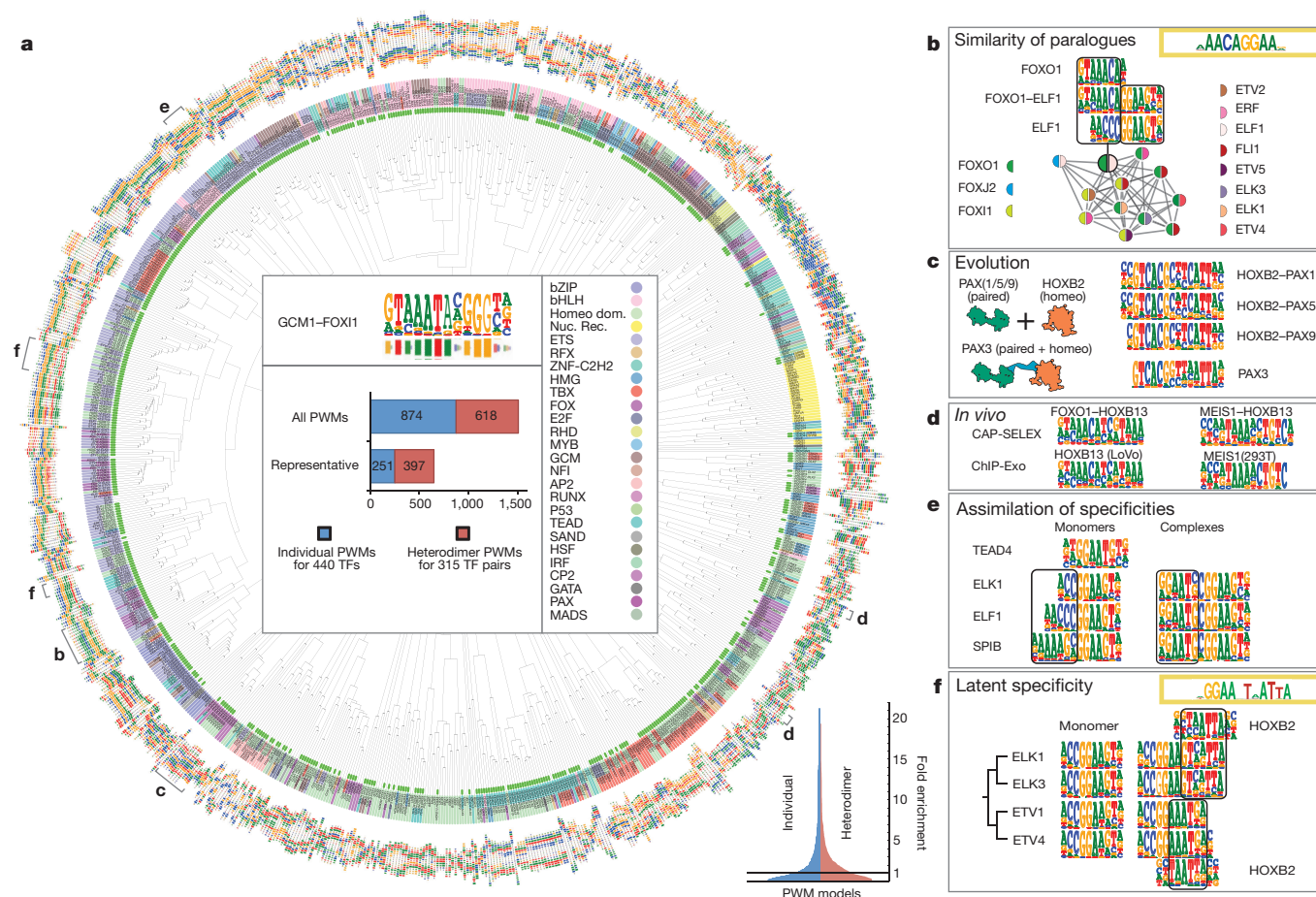
palindromic recognition sequences can only have two or one orientations, respectively. These results indicate that TF-TF cooperativity is widespread and not just mediated by the highly specific protein-protein interactions observed in previously described canonical heterodimers.



**Figure 2 | Overlapping composite TF motifs with novel specificity.** **a**, An example of a TF pair binding to an overlapping composite site. Top, a composite GCM1-ELK1 logo aligned to the individual logos. Middle, DNA-protein contacts for GCM1 (purple) and ELK1 (light blue) in the composite site, predicted based on GCM1-DNA and ELK1-DNA structures<sup>29,30</sup>. Dots indicate the number of hydrogen bonds between the TF and DNA backbone (ovals) and bases (rectangles) occurring via major (filled circle) and minor (open circle) grooves. Bottom, a schematic model of GCM1-ELK1 heterodimer. DNA is shown as idealized B-DNA. **b**, TFs prefer to bind to sites where their core motifs are closely spaced. Histogram of gaps observed between all full width motifs (core plus flank) and core motifs. Gap widths were counted for all TF pairs identified in this study for which structural data was available. Examples of calculating distance using full PWMs (above  $x$  axis) and core motifs (below  $x$  axis) are also shown.

Some TF pairs displayed strong orientation and spacing preferences, without major changes in either motif (Fig. 1b, c). However, in a large number of cases (207), the specificity of the pair of TFs differed markedly from that expected from the individual motifs (Fig. 2a). These differences were observed when the two TFs were close to each other. To understand the mechanism of the altered specificity, we analysed available structures for the studied TFs and their paralogues (Supplementary Data Set 2). Based on the analysis, 95% of the complexes are consistent with either a completely DNA-mediated mechanism, or a DNA-facilitated mechanism, where the interaction between the proteins is scaffolded by DNA and limited to few amino-acid contacts; only 5%





**Figure 3 | All identified TF–TF interactions.** **a**, PWM motif similarities between the heterodimer motifs (green bars) and monomeric and homodimeric representative motifs from ref. 8. Barcode logos for each factor are shown, and background colour of name indicates TF structural family. Center of dendrogram shows comparison of sequence and barcode logos, the colour key, and the number of all PWMs and representative PWMs. Inset (bottom right), fold enrichment of matches of the motifs in known TF clusters from human colon cancer cells<sup>19</sup>. **b**, A network representation of the very similar heterodimeric sites formed between multiple FOX and ETS proteins. Note that a similar site is recognized when

of the complexes appeared to form extensive protein–protein interaction surfaces (Extended Data Fig. 5). We next generated position weight matrices (PWMs) that included information about hydrogen bond contacts between the TF amino acids, and DNA bases and backbone (Supplementary Table 3). Alignment of pairs of such contact-annotated PWMs to their respective composite models revealed that the changes in binding specificity mostly affected base positions that are recognized by the TFs via contacts to the DNA backbone. In contrast, ‘core’ bases directly read via hydrogen bonds were rarely affected (Fig. 2b and Extended Data Fig. 6).

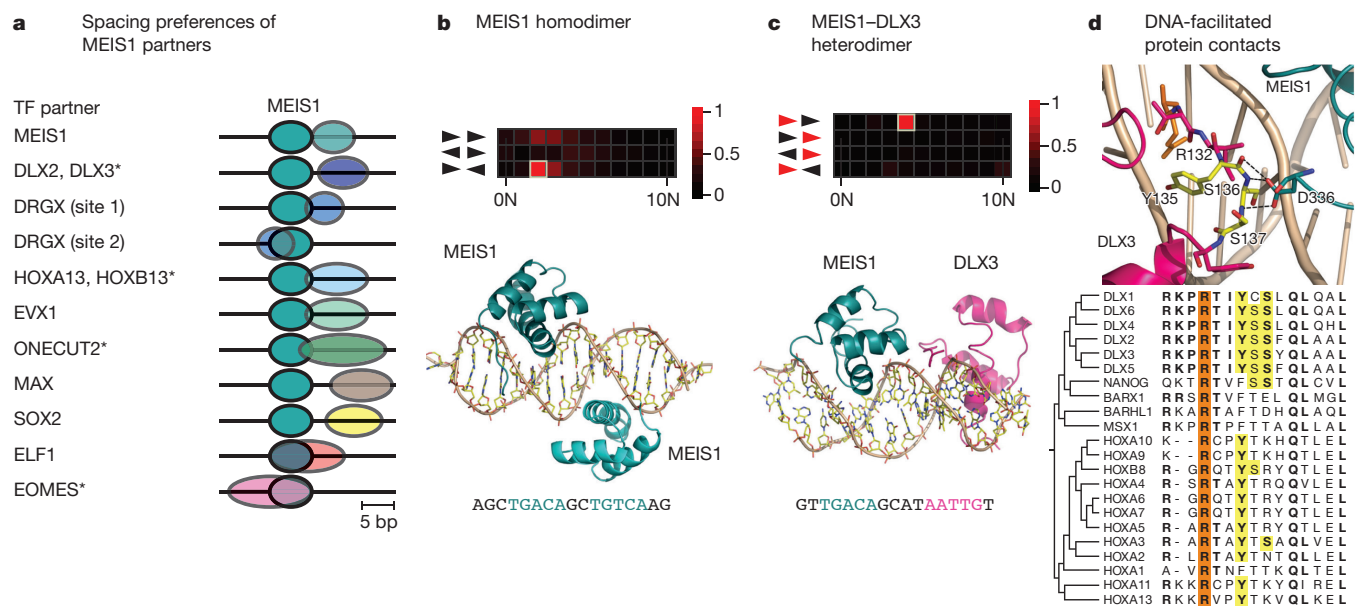
In total, we recovered 618 PWM models describing the specificities of the TF–TF pairs. To globally analyse the collection, we identified distinctly different ‘representative’ motifs from a combination of our data set and previous high-throughput SELEX (HT-SELEX) data<sup>8</sup>. Out of all representative motifs, 61% were TF pair motifs identified in this study (Fig. 3 and Extended Data Fig. 7), suggesting that a very large fraction of TF specificity space is defined by TF heterodimers. A dendrogram displaying the similarity of the heterodimeric motifs revealed that paralogous proteins often shared the same partners, and bound to similar heterodimeric motifs (Fig. 3a). A total of 63 of such motif groups were identified, representing 239 TF pairs. For example, many ETS factors formed complexes with forkhead proteins and with the posterior

the FOX protein is either used as TF1 (FOXO1, FOXJ2) or TF2 (FOXI1). Similar conserved motif is also shown<sup>20</sup>. **c**, HOXB2–PAX1, HOXB2–PAX5 and HOXB2–PAX9 heterodimer sites are similar to a site for PAX3, which contains both Pax- and homeodomains. **d**, FOXO1–HOXB13 and MEIS1–HOXB13 heterodimers validated by ChIP-exo. **e**, Binding of TEAD4 together with the indicated ETS TFs makes their divergent flanking recognition sequences (left box) more similar to each other (right box). **f**, HOXB2 reveals latent specificity of the TFs indicated. Inset shows conserved genomic motif<sup>20</sup> that is similar to the ELK1–HOXB2 motif.

homeodomain TFs HOXD12 and HOXB13, binding to highly similar composite sites (Fig. 3b and Supplementary Table 2). Furthermore, PAX proteins containing only a paired domain interacted with homeodomain-containing partners, binding to sites that were similar to those recognized by PAX proteins that include both paired domains and homeodomains. This suggests that this site predates the joining of the paired domain and homeodomain to the same gene (Fig. 3c).

HOXB13 also formed complexes with forkhead and MEIS proteins. The motifs recovered using CAP-SELEX were also enriched by HOXB13 ChIP-exo<sup>18</sup>. The preferred dimer partner of HOXB13 was cell-line specific, suggesting that TFs dimerize with different proteins in different cell types (Fig. 3d). The inclusion of HOXB13 and MEIS1 dimer motifs also improved prediction of the corresponding ChIP-seq peak positions (Extended Data Fig. 5 and Supplementary Table 4).

The novel motifs were enriched in ChIP-seq-identified TF cluster sequences<sup>19</sup> (Fig. 3a); a larger fraction (52%) of the novel motifs were enriched compared to the monomer motifs (34%, Supplementary Tables 2 and 4). Furthermore, comparison to motifs discovered *de novo*<sup>20</sup> and our independent analysis revealed that many (24%) of the motifs were enriched in mammalian conserved sequences (Supplementary Table 2 and Extended Data Figs 5 and 8). A total of 390 of 618 motifs were found enriched in human TF clusters and/or mammalian conserved sequences. Both of these methods have a relatively



**Figure 4 | Structural validation of TF–TF interactions.** **a**, Positions of MEIS1 partner TFs in relation to the 7 bp MEIS1 motif<sup>8</sup> (cyan, orientation NTGACAN). Note that the partners bind to different positions, spanning a 26-bp region. Modelling suggests that in all pairs except MEIS1–ELF1, the proteins do not interact extensively (Supplementary Data Set 2). Asterisks indicate that the corresponding genomic motif matches are conserved in mammals (see Supplementary Table 2). **b**, Structure of MEIS1–MEIS1–DNA complex. Top, Heatmap based on HT-SELEX data<sup>8</sup> showing occurrence of MEIS1 subsequence TGACA in the orientations indicated (arrowheads, scale divided by highest count). The preferred spacing and orientation is indicated by yellow outline. Bottom, structure of two MEIS1 proteins

bound to the preferred site. TGACA and its reverse complement sequence are in cyan. **c**, Structure of MEIS1–DLX3–DNA complex. Top, heatmap showing occurrence of MEIS1 and DLX3 5-mer subsequences TGACA (black arrowhead) and AATTG (red arrowhead), respectively. Bottom, crystal structure of MEIS1 and DLX3 bound to the preferred site. Note the narrowing of the DNA minor groove between the two proteins. **d**, DLX3 Arg132 (orange) inserts into the minor groove of DNA, and positions two adjacent serines and a tyrosine (yellow) so that an aspartate from MEIS1 hydrogen bonds (dotted lines) with DLX3 peptide backbone. Bottom, conservation of the residues in homeodomain proteins. Residues conserved in all human DLX proteins are in bold.

high false-negative rate, due to differences in dimers in different cell types, and the requirement that >50 motif occurrences are conserved to reach statistical significance. These results indicate that motifs identified in this study are biologically relevant.

Heterodimeric partners could also mask differences in binding specificities of individual TFs. For example, class I, II and III ETS factors ELK1, ELF1 and SPIB, respectively, prefer different 5' flank sequences<sup>8,21</sup>, but this difference is effectively masked by TEAD4 (Fig. 3e). This effect was rare; only two other similar cases were identified (Supplementary Data Set 1). Conversely, partners could be identified that revealed 'latent specificity'<sup>14</sup> of TFs, defined as binding of TFs to different heterodimeric sites, even when their primary specificities are indistinguishable. For example, ETV1, ETV4, ELK1 and ELK3 bind to similar monomeric sites, and also bound to similar heterodimeric sites with GCM1 (Supplementary Table 2). However, with HOXB2, ETV1 and ETV4 bound to one type of site, and ELK1 and ELK3 to another site (Fig. 3f). The ETVs are more closely related to each other than to the ELKs, suggesting that latent specificity evolves faster than primary specificity. Four other similar cases were identified (Supplementary Data Set 1).

To analyse the mechanisms of cooperativity, we studied all identified dimers that included the TALE-class homeodomain MEIS1. Twelve TFs from six TF families bound to diverse but specific positions at either side of MEIS1 (Fig. 4a). To understand the basis of such interactions, we solved the structure of MEIS1 bound to DNA alone, as a homodimer, and as a heterodimer with DLX3 using X-ray diffraction (3.5 Å, 1.6 Å and 3.5 Å resolutions, respectively). In the homodimer structure, the two monomers are on opposite sides of DNA and do not contact each other, indicating that the observed cooperativity is entirely mediated by DNA (Fig. 4b).

Interaction between MEIS1 and DLX3 in CAP-SELEX is much stronger than that observed for the MEIS1–MEIS1 dimer<sup>8</sup>. The proteins

interact, but the contact surface is very small, covering only 2.0% of the solvent accessible surface of the dimer. However, the DNA between the proteins is significantly deformed, narrowing the minor groove (Fig. 4c and Extended Data Fig. 9). This facilitates interaction between the proteins, as Arg132 of DLX3 inserts into the minor groove, positioning the conserved amino acids Tyr135, Ser136 and Ser137 in such a way that the peptide backbone makes three hydrogen bond contacts with Asp336 of MEIS1 (Fig. 4d).

In summary, our sampling of a large number of TF–TF interactions revealed a much greater number of interactions than previously reported. Many novel DNA motifs were enriched in ChIP-seq TF clusters and conserved in mammalian genomes. Based on the fraction of pairs tested, we estimate that ~25,000 distinct TF pair specificities contribute to protein–DNA interactions in cells (Supplementary Table 4). The frequent interactions between TFs, together with nucleosome-mediated cooperativity<sup>22,23</sup> are consistent with the observation that TF binding in cells occurs in dense clusters. The clusters are also likely to be stabilized by TF co-factors, and complexes such as Mediator, cohesin and RNA polymerase II<sup>19,24,25</sup>. Such higher-order complexes, and complexes between TFs formed by domains that were not included in our constructs (Supplementary Table 1) are likely to further contribute to the cooperativity of TF binding *in vivo*.

Most of the observed interactions involve close packing of the individual TF's core motifs, and overlap between the motif flanks. The sequence between the core motifs commonly differs from that expected from the individual motifs. The composite sites would often be recognized by the individual TFs, but with relatively low affinity. Our findings are thus consistent with the general low affinity of sites bound *in vivo* in ChIP-seq experiments, and the fact that the conservation pattern of many regulatory elements extends beyond known TF binding sites. Taken together, our results show that cooperativity is an inherent feature of TF–DNA binding, and that DNA itself functions as an active

interacting partner, commonly facilitating the interactions between a wide range of TFs from diverse structural families.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 23 January; accepted 24 August 2015.**

**Published online 9 November 2015.**

- Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nature Rev. Genet.* **15**, 272–286 (2014).
- Levo, M. & Segal, E. In pursuit of design principles of regulatory sequences. *Nature Rev. Genet.* **15**, 453–468 (2014).
- Slattery, M. *et al.* Absence of a simple code: how transcription factors read the genome. *Trends Biochem. Sci.* **39**, 381–399 (2014).
- Rodda, D. J. *et al.* Transcriptional regulation of *Nanog* by OCT4 and SOX2. *J. Biol. Chem.* **280**, 24731–24737 (2005).
- Panne, D., Maniatis, T. & Harrison, S. C. An atomic model of the interferon- $\beta$  enhanceosome. *Cell* **129**, 1111–1123 (2007).
- De Val, S. *et al.* Combinatorial regulation of endothelial gene expression by Ets and Forkhead transcription factors. *Cell* **135**, 1053–1064 (2008).
- Badis, G. *et al.* Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720–1723 (2009).
- Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell* **152**, 327–339 (2013).
- Vaquerezas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nature Rev. Genet.* **10**, 252–263 (2009).
- Najafabadi, H. S. *et al.* C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nature Biotechnol.* **33**, 555–562 (2015).
- Emery, P. *et al.* A consensus motif in the RFX DNA binding domain and binding domain mutants with altered specificity. *Mol. Cell. Biol.* **16**, 4486–4494 (1996).
- Kurokawa, R. *et al.* Differential orientations of the DNA-binding domain and carboxy-terminal dimerization interface regulate binding site selection by nuclear receptor heterodimers. *Genes Dev.* **7**, 1423–1435 (1993).
- Mohibullah, N., Donner, A., Ippolito, J. A. & Williams, T. SELEX and missing phosphate contact analyses reveal flexibility within the AP-2 $\alpha$  protein: DNA binding complex. *Nucleic Acids Res.* **27**, 2760–2769, (1999).
- Slattery, M. *et al.* Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell* **147**, 1270–1282 (2011).
- Grove, C. A. *et al.* A multiparameter network reveals extensive divergence between *C. elegans* bHLH transcription factors. *Cell* **138**, 314–327 (2009).
- Nitta, K. R. *et al.* Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *eLife* **4**, e04837 (2015).
- Kim, S. *et al.* Probing allostery through DNA. *Science* **339**, 816–819 (2013).
- Rhee, H. S. & Pugh, B. F. Comprehensive genome-wide protein–DNA interactions detected at single-nucleotide resolution. *Cell* **147**, 1408–1419 (2011).
- Yan, J. *et al.* Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. *Cell* **154**, 801–813 (2013).
- Guturu, H., Doxey, A. C., Wenger, A. M. & Bejerano, G. Structure-aided prediction of mammalian transcription factor complexes in conserved non-coding elements. *Phil. Trans. R. Soc. Lond. B* **368**, 20130029 (2013).
- Wei, G. H. *et al.* Genome-wide analysis of ETS-family DNA-binding *in vitro* and *in vivo*. *EMBO J.* **29**, 2147–2160 (2010).
- Mirny, L. A. Nucleosome-mediated cooperativity between transcription factors. *Proc. Natl Acad. Sci. USA* **107**, 22534–22539 (2010).
- Wasson, T. & Hartemink, A. J. An ensemble model of competitive multi-factor binding of the genome. *Genome Res.* **19**, 2101–2112 (2009).
- Poss, Z. C., Ebmeier, C. C. & Taatjes, D. J. The Mediator complex and transcription regulation. *Crit. Rev. Biochem. Mol. Biol.* **48**, 575–608 (2013).
- Kagey, M. H. *et al.* Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467**, 430–435 (2010).
- Nishizawa, M. & Nagata, S. cDNA clones encoding leucine-zipper proteins which interact with G-CSF gene promoter element 1-binding protein. *FEBS Lett.* **299**, 36–38 (1992).
- Shen, W. F. *et al.* AbdB-like Hox proteins stabilize DNA binding by the Meis1 homeodomain proteins. *Mol. Cell. Biol.* **17**, 6448–6458 (1997).
- Williams, D. C., Jr, Cai, M. & Clore, G. M. Molecular basis for synergistic transcriptional activation by Oct1 and Sox2 revealed from the solution structure of the 42-kDa Oct1·Sox2·Hoxb1–DNA ternary transcription factor complex. *J. Biol. Chem.* **279**, 1449–1457 (2004).
- Cohen, S. X. *et al.* Structure of the GCM domain–DNA complex: a DNA-binding domain with a novel fold and mode of target site recognition. *EMBO J.* **22**, 1835–1845 (2003).
- Mo, Y., Vaessen, B., Johnston, K. & Marmorstein, R. Structure of the Elk-1–DNA complex reveals how DNA-distal residues affect ETS domain recognition of DNA. *Nature Struct. Mol. Biol.* **7**, 292–297 (2000).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank J. Yan, B. Schmierer, E. Kaasinen, C. Daub, E. Haapaniemi and Å. Kolterud for their review of the manuscript, the Karolinska Institutet protein science facility for protein purification, and S. Augsten, L. Hu and A. Zetterlund for technical assistance. This work was supported by Finnish Academy CoE in Cancer Genetics, Center for Innovative Medicine, Knut and Alice Wallenberg and Göran Gustafsson Foundations and Vetenskapsrådet.

**Author Contributions** A.J. and J.T. designed the experiments. A.J. and Y.Y. performed CAP-SELEX, K.D. performed ChIP-exo, and M.T. the sequencing analyses. A.J., K.R.N., T.K., M.E. and J.T. wrote computer programs for the analyses. E.M. and A.P. performed X-ray crystallography, and E.M. solved the structures. A.J., K.R.N. and E.M. prepared illustrations and A.J., Y.Y. and J.T. wrote the article. All authors contributed to data analysis and reviewed the manuscript.

**Additional Information** Sequencing reads are deposited to European Nucleotide Archive (accession PRJEB7934). The atomic coordinates and diffraction data are deposited to Protein Data Bank (accession 4XRM, 5BNG and 4XRS). All computer programs and scripts used are either published or available upon request. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.T. ([jussi.taipale@ki.se](mailto:jussi.taipale@ki.se)).



## METHODS

**Sequencing and data analysis.** Unselected initial libraries and products of the third selection cycle were purified using a PCR-purification kit (Qiagen) and sequenced using Illumina HiSeq 2000 (multiplexed as in ref. 31; 55 bp single-read length). Raw sequence reads were demultiplexed, and initial quality control was performed using IniMotif<sup>31</sup>. Sequencing depth was set in such a way that on average each experiment would result in 250,000 sequence reads. Based on previous enrichment ratios, this should lead to more than 1,000 highly enriched seed subsequences to be detected (count statistics; Poisson distribution, 3.16% standard deviation; expected background for 10 bp seed 15 counts,  $P$  value  $1.26 \times 10^{-273}$  using winflat, expected 15, observed  $\geq 1,000$ ; Bonferroni corrected  $P$  value  $< 1.32 \times 10^{-267}$ ). Average background-corrected count for the seed match at the indicated multinomial setting was 3,295 (Supplementary Table 2). All sequence data has been deposited to ENA (European Nucleotide Archive) under accession number PRJEB7934.

To identify overlapping composite sites, we used the AUTOSEED tool described in Nitta *et al.*<sup>16</sup>. This tool is based on identification of gapped and ungapped subsequences that represent a local maxima within a Huddinge distance<sup>16</sup> of 1; that is, they are more enriched than all subsequences that align with them with  $k-1$  perfectly matching bases, where  $k$  is the length of the ungapped subsequences.

**Cell culture and ChIP-exo.** LoVo (source: ATCC, cat. no. CCL229TM), 293FT (source: Thermo Fisher Scientific, cat. no. R700-07) and GP5d (source: ECACC, cat. no. 95090715) cells were cultured in DMEM supplemented with 10% fetal bovine serum (FBS) and antibiotics. Cells were obtained directly from the indicated source, and tested and found negative for mycoplasma contamination by immunofluorescence analysis after staining with 3.3  $\mu\text{g ml}^{-1}$  bisBenzimide H 33342 trihydrochloride (Sigma cat. no. B2261). All antibodies used in ChIP-exo experiments were ChIP-grade. In each experiment a non-specific IgG was used as a control. ChIP-exo was performed essentially as described in Rhee and Pugh<sup>18</sup> with modifications from ref. 32 using antibodies for HOXB13, MEIS1 and rabbit IgGs (Santa Cruz Biotechnology and Abcam cat. numbers sc-66923X, ab19867, and sc-2027, respectively). See Supplementary Table 1 for the sequences of the Illumina sequencing adapters. Sequence reads were mapped to the human reference genome (hg18), using bwa with default parameters. For peak-calling, we used GEM<sup>33</sup> with default parameters, and the genome size set to 2,700,000,000.

**Construct design.** TFs were selected to cover different structural classes and individual binding specificities. Thus, in the set, small TF families such as TEA and GCM are relatively overrepresented, and large families such as C<sub>2</sub>H<sub>2</sub> zinc fingers and canonical homeodomains are underrepresented. The expression constructs contained the DNA-binding domain, and known dimerization and protein-protein interaction domains for TF families where such domains are known to be required for DNA binding. These included, for example leucine zipper domains of bZIP and bHLH proteins, pointed-domains of ETS factors, dimerization domains of nuclear receptors as well as short motifs such as 'YPWM' of the anterior homeodomains that are known to be involved in protein-protein interactions<sup>34</sup> (see Supplementary Table 1 for full sequences and removed and retained domains). Flanking sequences of 15 amino acids were also included on both sides to allow folding of the known protein domains, and to retain amino acids that are located close to DNA and could mediate interactions between closely packed TFs. We have previously shown that such constructs recover accurate binding specificities and homodimeric interactions by analysis of 125 pairs of such constructs and the corresponding full-length TF proteins<sup>8,16</sup>.

**Protein expression, purification and activity testing.** Bacterial protein expression Gateway recipient vectors that incorporated a N-terminal Thioredoxin-6×His tag, with either a C-terminal streptavidin binding peptide (SBP) or 3×Flag tag were constructed using pETG20A-plasmid as a backbone. Inserts for protein expression were derived either by gene synthesis (Genscript; see Supplementary Table 1 for protein sequences and domains), or from previously published Gateway donor clones<sup>8</sup>. All proteins were expressed in the Rosetta 2(DE3)pLysS *E. coli* strain as fusion proteins using the auto-induction protocol described in Jolma *et al.*<sup>8</sup>. Proteins were purified using nickel affinity purification (GE Nickel sepharose Fast-Flow6, GE, Sweden) and stored at  $-20^\circ\text{C}$  in 50% glycerol, 150 mM NaCl, 250 mM imidazole, 15 mM Tris-Cl, pH 7.5.

Protein expression and purification from *E. coli* cells was performed as described in Jolma *et al.*<sup>8</sup>. Briefly, the expression system used Rosetta 2(DE3)pLysS strain of *E. coli* (Millipore) cultured in ZYP5052 autoinduction media, where the expression of proteins is induced upon consumption of the preferred carbon source (glucose). Transformed cells were first cultured in deep well 96-well plate (Thermo, AB0661) wells in TB-medium at  $37^\circ\text{C}$  for overnight and then transferred to the auto-induction medium (1:40 dilution, see Vincentelli *et al.*<sup>35</sup>). When the cell density was between 2.0 and 3.0 optical density at 600 nm, the temperature was lowered to  $17^\circ\text{C}$ , and the culture continued for 40 h. The cells were harvested by centrifugation (4,000 rpm for 15 min), and lysed by incubation with buffer A (300 mM NaCl

in 50 mM Tris-Cl, pH 7.5) containing 10 mM imidazole, 0.5 mg ml<sup>-1</sup> lysozyme (Sigma) and 1 mM PMSF (Sigma). The lysis was completed by a freeze-thaw cycle. DNase I and MgSO<sub>4</sub> were added to the thawed lysate at 10  $\mu\text{g ml}^{-1}$  and 1 mM final concentration, respectively, and the lysates incubated with Ni-Sepharose 6 Fast Flow resin (GE Healthcare) and shaken for 45 min. The lysate was then transferred to a filter plate (Nunc, 278011, 20  $\mu\text{m}$  pore size), and the beads washed two times each with 600  $\mu\text{l}$  of 10 mM and 50 mM imidazole in buffer A using a vacuum manifold. The bound proteins were eluted from resin using 500 mM imidazole in buffer A. The expression of the purified proteins were checked by UV absorbance at 280 nm and SDS-PAGE electrophoresis (E-PAGE protein gels, Invitrogen) and Coomassie brilliant blue staining. 50% glycerol was added to the proteins before storage at  $-20^\circ\text{C}$ .

In most cases the activity of the proteins was assessed by HT-SELEX<sup>8,31</sup>, and proteins that robustly enriched expected sequences were included in the CAP-SELEX process. As some TFs are only expected to bind to DNA as a heterodimer, we also included in CAP-SELEX some proteins that did not robustly enrich sequences in HT-SELEX. These included the known obligate heterodimers MYC, PBX1, PBX2 and PBX4. All included proteins are indicated in Supplementary Table 1. The HT-SELEX analysis yielded expected binding sites for most individual TFs, and in addition resulted in identification of novel motifs for TFs (see Supplementary Table 2). HT-SELEX was also used to validate some of the CAP-SELEX results with full length TFs. In these cases the pair of proteins were mixed together in an ~1:1 ratio before the further steps of the HT-SELEX (see Supplementary Table 1 for the details of the clones).

In most cases, a scaled-up culture (50 to 100 ml) was used to express more of the TF constructs for CAP-SELEX. The protocol used was similar to that used for the deep-well plate cultures, except that proportionally larger lysis and wash volumes were used, and that 1 ml Ni-Sepharose 6 Fast Flow gravity columns (GE Healthcare) were used as the affinity matrix.

**CAP-SELEX assay.** Previous systematic efforts that have focused on heterodimerization between TFs have generally studied proteins that dimerize in the absence of DNA<sup>36-38</sup>. However, some cases of TFs that cannot dimerize in the absence of DNA, or only interact with each other indirectly through DNA have been described in the literature<sup>39,40</sup>. The CAP-SELEX process can capture both types of interactions, and is based on a combination of HT-SELEX<sup>8,31</sup> with tandem-affinity purification<sup>41</sup>. In this assay, a pair of TFs tagged with different affinity tags, SBP and 3×Flag, and a double stranded DNA ligand that contains a 40-bp randomized region are mixed together in individual wells of a 96-well plate in a buffer that mimics biological conditions in the nucleus<sup>42</sup>, and the mixture is incubated for 30 min, after which the bound dsDNA ligands are separated from free ligands through consecutive affinity purification by first the SBP and then the 3×Flag tagged protein using affinity beads and an automated plate washer (Fig. 1a). Bound DNA is then amplified by PCR and sequenced, and the selection process repeated three times. Binding of the TFs is then revealed by enrichment of characteristic subsequences (see below).

The 40-bp random window corresponds to almost four complete turns of the DNA helix, allowing detection of interactions between two TFs that exclusively occupy 9 bp of sequence (see ref. 20) over two full helical turns. As with our previous HT-SELEX platform, the purified ligands are barcoded and directly compatible with multiplexed Illumina sequencing (for selection ligand sequences, please see Supplementary Table 1).

Of the proteins tested here, most (87%) were functionally validated by HT-SELEX (see Supplementary Table 1 for details). The low activity of some HT-SELEX validated proteins (Supplementary Table 1) was probably due to the fact that CAP-SELEX involves two consecutive affinity purifications, and is therefore more stringent than HT-SELEX.

For CAP-SELEX, 10–200 ng (see Supplementary Table 1) purified Flag- and SBP-tagged proteins were diluted into 25  $\mu\text{l}$  volume of binding buffer (140 mM KCl, 5 mM NaCl, 2 mM MgSO<sub>4</sub>, 3  $\mu\text{M}$  ZnSO<sub>4</sub>, 100  $\mu\text{M}$  EGTA, 1 mM K<sub>2</sub>HPO<sub>4</sub>, in 20 mM HEPES, pH 7.0) containing approximately 10  $\mu\text{mol}$  DNA selection ligands, and incubated for 20 min at room temperature. Subsequently, 0.2% BSA, 0.1% Tween 20 pre-blocked Streptavidin-coated magnetic Sepharose beads (1.25  $\mu\text{l}$ ; GE Healthcare Streptavidin Mag Sepharose) in two volumes of binding buffer were added, and the mixture incubated at room temperature for 2 h with vigorous shaking (800 r.p.m.; Edmund Bühler TiMix shaker). The beads were subsequently washed 5 times with binding buffer, using BioTek 405 CW plate washer with a magnetic platform. The protein-DNA complexes were eluted from the beads using 50  $\mu\text{l}$  of 10 mM biotin (Sigma) in binding buffer. The eluate was transferred to a fresh plate containing M2 anti-Flag magnetic beads (1.25  $\mu\text{l}$ ; Sigma) in 50  $\mu\text{l}$  of binding buffer, and shaken at 800 r.p.m. for 20 min at room temperature. The beads were washed ten times with binding buffer, suspended in 0.1% Tween 20, 0.5 mM EDTA, 10 mM Tris-Cl, pH 8.0, and transferred to a PCR plate. DNA was then eluted from the beads by

incubation at 95°C for 10 min. A 9 µl aliquot of the bead suspension was transferred to a new PCR plate, and the DNA amplified by PCR (65°C for 10 s, 72°C for 36 s, 97°C for 15 s for annealing, elongation and denaturation, respectively, for 25 cycles). A separate aliquot was analysed by qPCR (Roche LightCycler 480) to monitor progress of the experiment. Amplified selection ligands were then subjected to sequencing and new cycles of CAP-SELEX (up to 3 cycles total).

The input libraries and libraries selected for three cycles were then sequenced and analysed (see below). Cooperative complexes were initially detected from 5–12-bp long primary and secondary binding models generated by the previously described SELEX data-analysis tool IniMotif<sup>31</sup>. Initial results validated the method through confirmation, characterization and refinement of the binding specificity models for 12 previously known heterodimers (Fig. 1a and Extended Data Fig. 1). **Generation of PWMs.** To detect heterodimeric sites that can occur in many different orientations and spacings, we used the *de novo* motif discovery algorithm Autoseed<sup>16</sup>. Autoseed finds gapped and ungapped subsequences that represent local maxima, that is, are more enriched than closely related subsequences. Control experiments established that such preferences were not observed in the input libraries.

Based on analysis of CAP-SELEX cycle 3, the sequencing was then extended to cover products from earlier selection cycles for samples that showed enrichment of motifs that were similar to the expected TF2 motif. We have previously shown that analysing SELEX data from early cycles allows recovery of low-affinity sites, and results in motifs that are very similar to those determined using competition assays<sup>8,16</sup>.

For identification of co-operative sites where the individual TF motifs were spaced farther apart, we defined representative 6-mer sequences for each TF. For each experiment, we then identified cases where both representative 6-mers were found in the same sequence reads, and from these reads, counted the occurrence of each spacing and orientation combination. The expected distribution of spacings without any preferences is non-uniform due to the limited size of the randomized region but cannot explain the local maxima observed in the data (spacings that are preferred compared to both shorter and longer spacings). Given the size of both occupied sites  $m$  and a fixed spacing  $n$  between them, the number of ways the sites can be placed into the randomized region of size  $l$  is  $c_n = l - 2m - n + 1$  ( $0 \leq n \leq l - 2m$ ), or twice that if the order of sites is taken into account. As the expected count of spacing  $n$  is directly proportional to  $c_n$ , the count is a linearly decreasing function of  $n$  and thus has no maxima except the one at the boundary  $n = 0$ . Moreover, the expected relative difference of counts between consecutive spacings is  $(c_n - 1 - c_{n+1})/c_n = 1/c_n$  indicating that the expected differences are small in the range where spacing preferences are observed.

To display the preferred spacings and orientations of the two TFs, subsequences containing both consensus 6-mer sequences of the motifs with variable-length gap between them were counted, and the counts represented as a heatmap (for example, Fig. 1). For each orientation of the 6-mers, we identified the spacing with maximum counts. That spacing was considered to be preferred if the sum of the counts of it and its two neighbours was higher than 30% of the total count for all spacings, and less than 20% of the reads counted were derived from a single non-unique read. If one preferred case was identified for a pair of 6-mers, we then determined whether other spacings and orientations were also enriched. Up to five spacings and orientations were considered to be preferred if their respective counts were higher than 50% of the maximum count after mean normalization of all counts. Cases where both TF1 and TF2 6-mers were detected robustly, but no preferred orientation and spacing was detected were classified as 'weak or no co-operativity'. In case of experiments performed several times, interactions were called if they passed the thresholds in at least one case. In each case, control unselected ligands were also sequenced, to ensure that the oligonucleotide synthesis resulted in even distribution of mononucleotides. In addition, several ligands were sequenced very deeply (>10 million reads) to ensure that 6-mer subsequences were distributed at a similar frequency along the 40-bp random sequence window (Extended Data Fig. 4). For assessment of reproducibility, see Extended Data Fig. 3.

Subsequently, the enriched subsequences were used as seeds to generate position weight matrix (PWM) models for the complexes. We generated PWM models as described in Jolma *et al.*<sup>8,31</sup>, using the seeds, selection cycles and multinomial setting indicated in Supplementary Table 2. The models were subsequently inspected to remove cases that could be also explained by homodimeric binding (for example, cases where two TFs with similar primary motifs were analysed). After identification of an enriched sequence, seed refinement was performed essentially as described in Jolma *et al.*<sup>8</sup>. In the majority of the cases, the PWM models were generated using selection cycle 3. The models were expert curated to separate different binding modes, and to remove cases where there was excessive positional bias or where the two proteins bound to very similar sites and thus we could not differentiate between homo- and heterodimeric binding. All seed, cycle and

multinomial settings used are indicated on Supplementary Table 2, and the sequence reads have been deposited to ENA under accession (PRJEB7934).

Hydrogen bond contacts between TF amino-acids and DNA bases were identified using the program CONTACT that is included in the CCP4 software suite<sup>43</sup>. Hydrogen bond information was added to PWMs to generate contact annotated PWMs (pfmc). PWMs were aligned to each other by minimizing the sum of individual base-to-base comparison scores calculated as follows: Max (information content for PWM1 position  $n$ , information content for PWM2 position  $m$ ) \* (Manhattan distance between base frequencies of PWM1 position  $n$  and PWM2 position  $m$ ). In regions where there was no overlap, the positions were compared to an equal frequency of all bases. Pairs of positions whose score was smaller than 0.25 are indicated by boxes in Supplementary Data Set 2. The same cut-off was used to count divergent base positions in Extended Data Fig. 6.

#### Enrichment of motifs at ChIP-seq TF clusters and prediction of ChIP-seq peaks.

Interactions between TFs appear to be important *in vivo*, as recent large scale genome-wide location analyses of TFs in cultured cells have revealed that in a given cell type, TFs bind only to a subset of their potential target sites, and that the occupied target sites are located in high-density clusters, where many different TFs colocalize within a few hundred bp long regions<sup>19,44,45</sup>. Many of the occupied sites do not contain high-affinity sites for the analysed TF, suggesting that co-operative interactions allow binding of TFs to low-affinity sites<sup>19,45,46</sup>. Formation of TF clusters is probably at least in part the result of competition between TFs and nucleosomes, which indirectly results in increased occupancy of TFs close to each other, even when the TFs do not have direct cooperative interactions<sup>22,23</sup>. Another mechanism that could contribute to TF cluster formation is direct co-operativity between TFs. To test this, we analysed enrichment of monomer and heterodimer PWM matches at human LoVo colon cancer TF clusters<sup>19</sup> (Fig. 3a inset, Supplementary Tables 2 and 4) as described in Yan *et al.*<sup>19</sup>, using a score cut-off for each motif that resulted in one match per 10 kb of genomic sequence.  $P$  values for the enrichment were calculated using winflat. Similarity of TF DBD amino-acid sequences was determined using PRANK<sup>47</sup>. Similarity of PWMs was determined using SSTAT<sup>48</sup> using the default parameters. Motif dendrogram was drawn by using Euclidean distance metric with average linkage with R package ape. Network shown in Fig. 3 was drawn based on the same distances. The dominating set of the models was generated essentially as described in Jolma *et al.*<sup>5</sup>. Briefly, we first generated a network where monomeric or homodimeric motifs from Jolma *et al.*<sup>8</sup>, and heterodimer motifs from this study were connected to each other if they were similar. To determine how many novel specificities were identified in our study, we used the minimum dominating set of this network to identify a set of motifs that represents distinctly different specificities.

Sequence and barcode-logos were generated as described in Nitta *et al.*<sup>16</sup>. In barcode logos shown in Fig. 3 and Extended Data Figs 3 and 7, four bars are drawn that represent the frequency of the bases for each base position. Width of each bar is proportional to the frequency of the corresponding base (range 0 to 1), and both height and colour intensity of all the bars at a given position are proportional to the frequency of the most common base at that position (range 0.25 to 1). In the DNA base letter PWM logos shown, the height of each letter is directly proportional to the frequency of the indicated base at the indicated position.

Analysis of error rates in prediction of ChIP-seq/exo peaks using the dimeric PWMs was performed using a random forest classifier. For the random forest classifier, the R package randomForest, version 4.6–6, was used (<http://cran.r-project.org/web/packages/randomForest/index.html>). The classifier was trained on TF motif matches to discern peak summit regions from close by non-peak summit genomic positions. Stated accuracy estimates are based on out-of-bag error estimates.

ChIP-seq binding site prediction error rate analysis was performed on data gathered from existing ChIP-seq experiments and from the HOXB13 ChIP-exo experiment from this study. The FASTA sequences of 1,001 bp genomic regions surrounding ChIP-exo peak summits for HOXB13 (from this work) and previously described ChIP-seq peak summits for ELF1 (ref. 21), ELK1 (ref. 49), HOXB13 (ref. 50) and two different MEIS1 experiments<sup>51,52</sup> were used as a positive set together with a negative set consisting of 1,001 bp FASTA sequences taken from 2,000 bp away from the peak summits. For each FASTA sequence in the collection, the score and relative position of the highest-scoring match to each motif in the PWM collection was recorded and used to train a randomForest classifier with 5,000 trees. To determine whether the dimer partners had predictive power, a classifier trained using the monomer motif of the relevant TF, and all its dimer motifs was compared to a classifier trained using the monomer motif and dimer motifs with the partner region of the motif reversed. Error rates were estimated using out-of-bag predictions.

**Analysis of conservation of motif matches.** To measure the conservation of genomic sites recognized by heterodimeric motifs we developed a procedure to test whether the heterodimeric motifs explained patterns of evolutionary conservation



observed in the human regulatory elements. To identify the potential conservation attributable to the specific TF–TF–DNA interactions described by the motifs, the genomic sites recognized by each heterodimeric motif were compared to sites recognized by artificial control motifs that represented different orientations of the two TFs but were otherwise comparable to the original motif, for example had the same width and information content. To obtain control motif sets containing the specificities of the individual TFs as embedded in the heterodimeric motifs, each heterodimeric motif was split into two partial motifs at all possible cut points with the restriction that both the ‘left’ and the ‘right’ sections were at least one-third of the width of the whole motif (rounded down). The control motifs were constructed by concatenating each of the partial motif pairs in all three alternative orientations (‘right’ followed by ‘left’, ‘left’ followed by the reverse complement of ‘right’, and the reverse complement of ‘left’ followed by ‘right’). For example, a motif of length fifteen was split after 5–10 bases resulting in 18 control motifs.

Sites recognized by a heterodimeric motif and its control motifs were searched from the human genome constrained elements<sup>53</sup> (SiPhy pi 12-mers with 10% false discovery rate in reference genome hg19, regions shorter than 50 bp or overlapping exons or repeats according to Ensembl version 70 were removed resulting in total 41 Mb of sequence) using the program MOODS<sup>54</sup> with a loose cut-off ( $P$  value  $< 10^{-3}$  with flat background distribution) to obtain a large excess of putative binding sites for each motif. All found sites were merged into one list and 10,000 non-overlapping highest affinity sites selected for conservation analysis regardless of the motif identity (heterodimer or control).

Whether the evolutionary conservation of the high affinity sites was explained by the motifs was tested using program SiPhy<sup>55</sup> (version 0.5, task 16, seedMinScore 0) and multiz100way multiple alignments<sup>56</sup> of 99 vertebrate species to human (downloaded from UCSC genome browser, version hg19). A site was marked as being conserved according to the motif if its SiPhy score was positive meaning that the aligned bases at the site were better explained by the motif than by a neutral evolutionary model (hg19.100way.phastCons.mod obtained from UCSC genome browser).

The hypothesis that the heterodimeric motif sites were more likely to be conserved according to the motif than the sites of its control motifs was tested against the null hypothesis that there was no association between site conservation and motif identity using Fisher’s exact test (one-sided). The  $P$  values given by the tests for individual heterodimeric motifs were corrected for multiple testing using Holm’s method. This procedure detected genomic conservation for 149 out of 618 motifs (24%) at family-wise error rate  $< 0.05$  (including the previously known ETV2–FOX11 motif<sup>20</sup> used as a positive control while developing the procedure).

**Protein purification, crystallization and data collection.** The DNA-binding domains of human MEIS1 (residues 277–339) and DLX3 (residues 122–193) were overexpressed as a thioredoxin-6His fusion protein in *E. coli* and isolated from the soluble cell lysate by affinity chromatography followed by gel-filtration chromatography as described in ref. 57. The DNA fragments used in crystallization were obtained from MWG as single strand oligonucleotides and annealed in 150 mM NaCl, 1 mM EDTA in 10 mM Tris-Cl, pH 7.5. The purified proteins were first mixed with solutions of annealed DNAs at a molar ratio of 1:1.2 and after incubation for 15–20 min on ice subjected to the crystallization trials. Crystallization experiments were carried out with an in-house developed crystal screening kit of different polyethylene glycols. The crystals of MEIS1–MEIS1–DNA complex were obtained in sitting drops at room temperature from 100 mM Tris-Cl (pH 8) solution containing 25.6% (w/v) PME (5000), 80 mM MgCl<sub>2</sub> and 10% PEG (400). The crystals of monodimer MEIS1–DNA complex were obtained from 100 mM HEPES (pH 7.09) solution containing 30% (w/v) PME (5000), 80 mM MgCl<sub>2</sub> and 10% PEG (400). Crystals of MEIS1–DLX3–DNA complex were obtained from 100 mM Tris-Cl (pH 7.5) containing 24% PEG (8000), 40 mM MgCl<sub>2</sub> and 5% butanol. All diffraction data for both complexes were collected at beam-line ID23-1 at the ESRF (Grenoble, France) using the reservoir solution as cryo-protectant. The data collection strategy was optimized with the program BEST<sup>58</sup>. The data were integrated with the program XDS<sup>59</sup> and scaled with SCALA<sup>60</sup>. Statistics of data collection are presented in Supplementary Table 5.

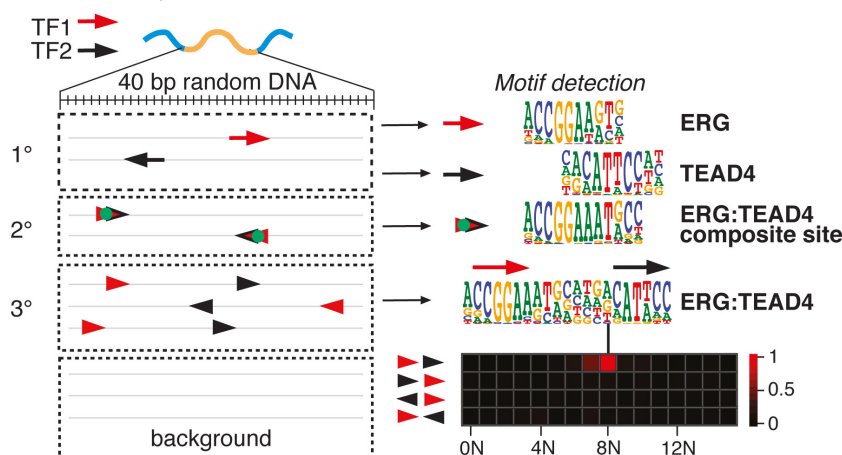
**Structure determination and refinement.** The structures of all three complexes were determined by molecular replacement using the program Phaser<sup>61</sup> in Phenix<sup>62</sup> with the structure of MEIS2 and DLX5 (PDB entries 3K2A and 2DJN, respectively) as a search model. The manual rebuilding of the model was done using COOT<sup>63</sup> combined with refinement with Phenix.refine using TLS option. The refinement statistics are presented in Supplementary Table 5. The atomic coordinates and diffraction data have been deposited to Protein Data Bank with the accession codes 4XRM, 5BNG and 4XRS, for MEIS1 homodimer, MEIS1 monomer and MEIS1–DLX3 heterodimer, respectively.

**Data reporting.** No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

1. Jolma, A. *et al.* Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.* **20**, 861–873 (2010).
2. Katainen, R. *et al.* CTCF/cohesin-binding sites are frequently mutated in cancer. *Nature Genetics* **47**, 818–821 (2015).
3. Guo, Y. *et al.* Discovering homotypic binding events at high spatial resolution. *Bioinformatics* **26**, 3028–3034 (2010).
4. Passner, J. M., Ryoo, H. D., Shen, L., Mann, R. S. & Aggarwal, A. K. Structure of a DNA-bound Ultrabithorax-Extradenticle homeodomain complex. *Nature* **397**, 714–719 (1999).
5. Vincentelli, R. *et al.* High-throughput protein expression screening and purification in *Escherichia coli*. *Methods* **55**, 65–72 (2011).
6. Keshava Prasad, T. S. *et al.* Human Protein Reference Database–2009 update. *Nucleic Acids Res.* **37**, D767–D772 (2009).
7. Ravasi, T. *et al.* An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* **140**, 744–752 (2010).
8. Newman, J. R. & Keating, A. E. Comprehensive identification of human bZIP interactions with coiled-coil arrays. *Science* **300**, 2097–2101 (2003).
9. Klemm, J. D. & Pabo, C. O. Oct-1 POU domain–DNA interactions: cooperative binding of isolated subdomains and effects of covalent linkage. *Genes Dev.* **10**, 27–36 (1996).
10. Panne, D., Maniatis, T. & Harrison, S. C. Crystal structure of ATF-2/c-Jun and IRF-3 bound to the interferon- $\beta$  enhancer. *EMBO J.* **23**, 4384–4393 (2004).
11. Rigaut, G. *et al.* A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnol.* **17**, 1030–1032 (1999).
12. Hallikainen, O. *et al.* Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell* **124**, 47–59 (2006).
13. Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallogr. D* **67**, 235–242 (2011).
14. Moorman, C. *et al.* Hotspots of transcription factor colocalization in the genome of *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA* **103**, 12027–12032 (2006).
15. Yip, K. Y. *et al.* Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol.* **13**, R48 (2012).
16. Meireles-Filho, A. C., Bardet, A. F., Yanez-Cuna, J. O., Stampfel, G. & Stark, A. cis-regulatory requirements for tissue-specific programs of the circadian clock. *Curr. Biol.* **24**, 1–10 (2014).
17. Löytynoja, A. & Goldman, N. An algorithm for progressive multiple alignment of sequences with insertions. *Proc. Natl Acad. Sci. USA* **102**, 10557–10562 (2005).
18. Pape, U. J., Rahmann, S. & Vingron, M. Natural similarity measures between position frequency matrices with an application to clustering. *Bioinformatics* **24**, 350–357 (2008).
19. Odrowaz, Z. & Sharrocks, A. D. The ETS transcription factors ELK1 and GABPA regulate different gene networks to control MCF10A breast epithelial cell migration. *PLoS ONE* **7**, e49892 (2012).
20. Huang, Q. *et al.* A prostate cancer susceptibility allele at 6q22 increases RFX6 expression by modulating HOXB13 chromatin binding. *Nature Genet.* **46**, 126–135, doi: 10.1038/ng.2862 (2014).
21. Huang, Y. *et al.* Identification and characterization of Hoxa9 binding sites in hematopoietic cells. *Blood* **119**, 388–398 (2012).
22. Penkov, D. *et al.* Analysis of the DNA-binding profile and function of TALE homeoproteins reveals their specialization and specific interactions with Hox genes/proteins. *Cell Reports* **3**, 1321–1333, (2013).
23. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).
24. Korhonen, J., Martinmaki, P., Pizzi, C., Rastas, P. & Ukkonen, E. MOODS: fast search for position weight matrix matches in DNA sequences. *Bioinformatics* **25**, 3181–3182 (2009).
25. Garber, M. *et al.* Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25**, i54–i62 (2009).
26. Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14**, 708–715 (2004).
27. Savitsky, P. *et al.* High-throughput production of human proteins for crystallization: the SGC experience. *J. Struct. Biol.* **172**, 3–13 (2010).
28. Bourenkov, G. P. & Popov, A. N. A quantitative approach to data-collection strategies. *Acta Crystallogr. D* **62**, 58–64 (2006).
29. Kabsch, W. Xds. *Acta Crystallogr. D* **66**, 125–132 (2010).
30. Collaborative Computational Project Number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
31. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
32. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
33. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).



64. Fitzsimmons, D. *et al.* Pax-5 (BSAP) recruits Ets proto-oncogene family proteins to form functional ternary complexes on a B-cell-specific promoter. *Genes Dev.* **10**, 2198–2211 (1996).
65. Kim, J. J. *et al.* Regulation of insulin-like growth factor binding protein-1 promoter activity by FKHR and HOXA10 in primate endometrial cells. *Biol. Reprod.* **68**, 24–30 (2003).
66. Vinson, C. R., Hai, T. & Boyd, S. M. Dimerization specificity of the leucine zipper-containing bZIP motif on DNA binding: prediction and rational design. *Genes Dev.* **7**, 1047–1058 (1993).
67. Williams, T. M., Williams, M. E. & Innis, J. W. Range of HOX/TALE superclass associations and protein domain requirements for HOXA13:MEIS interaction. *Dev. Biol.* **277**, 457–471 (2005).
68. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnol.* **30**, 271–277 (2012).
69. Raveh-Sadka, T. *et al.* Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. *Nature Genet.* **44**, 743–750 (2012).
70. Hochschild, A. & Ptashne, M. Cooperative binding of  $\lambda$  repressors to sites separated by integral turns of the DNA helix. *Cell* **44**, 681–687 (1986).
71. Moretti, R. *et al.* Targeted chemical wedges reveal the role of allosteric DNA modulation in protein–DNA assembly. *ACS Chem. Biol.* **3**, 220–229 (2008).
72. Aggarwal, A. K., Rodgers, D. W., Drott, M., Ptashne, M. & Harrison, S. C. Recognition of a DNA operator by the repressor of phage 434: a view at high resolution. *Science* **242**, 899–907 (1988).
73. Jordan, S. R. & Pabo, C. O. Structure of the lambda complex at 2.5 Å resolution: details of the repressor–operator interactions. *Science* **242**, 893–899 (1988).
74. Rohs, R. *et al.* Origins of specificity in protein–DNA recognition. *Annu. Rev. Biochem.* **79**, 233–269 (2010).

**a** CAP-SELEX data analysis**b** Comparison of CAP-SELEX PWMs to previous data

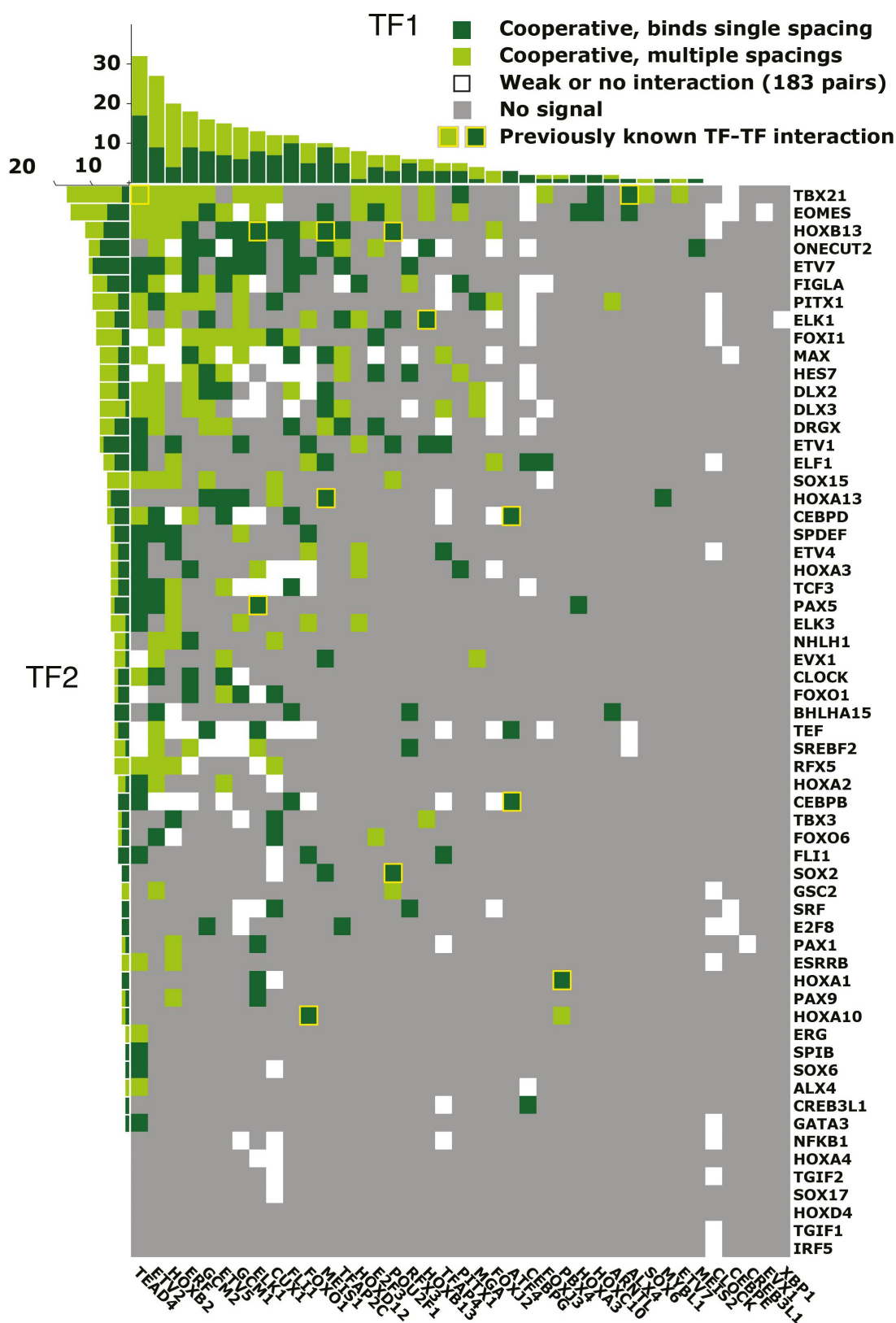
TF pair	CAP-SELEX model	Known consensus	Method	Ref.
ATF4:CEBPG*		CTGACGCAAT	EMSA	(26)
ATF4:CEBPB		CTGACGCAAT	EMSA	(66)
HOXB13:MEIS1*		GTCGTAAACTGTCA	EMSA	(27)
POU2F1:SOX2*		CATTAGCATGACAAAGACA	X-ray structure	(28)
ELK1:PAX5		GCCACTGGAGCCCATCTCCGGCA	EMSA	(64)

**c** CAP-SELEX PWMs for TF pairs with a known protein-protein interaction

TF pair	CAP-SELEX model	Method	
ALX4:TBX21		Mammalian two hybrid	(37)
TEAD4:TBX21		Mammalian two hybrid	(37)
POU2F1:HOXB13		Mammalian two hybrid	(37)
HOXB13:ELK1		Mammalian two hybrid	(37)
ATF4:CEBPD		Protein microarray	(38)
FOXO1:HOXA10		co-immunoprecipitation	(65)
HOXA13:MEIS1		Yeast two hybrid, co-immunoprecipitation	(67)

**Extended Data Figure 1 | CAP-SELEX data analysis and comparison to previous data.** **a**, Flowchart of CAP-SELEX data analysis. Left, a library of selection ligands with random sequences (yellow) is incubated with TFs. After CAP-SELEX, enriched individual TF motifs (1°; arrows) and composite motifs that are not simply combinations of the individual motifs (2°; green dots) are detected from the reads. To detect preferential spacings and orientations of the TF pair (3°), co-occurrence of the indicative 6-mer subsequences (arrowheads) are counted from the reads. The subsequences are then used to generate seeds for the PWM models (right). Heatmap (bottom right; scale divided by highest observed count) shows frequency of occurrence of the two 6-mers (CCGGAA, red arrowhead; CATTCC, black arrowhead) in all possible spacings (columns) and orientations (rows). Note that the 6-mer based approach cannot model the composite

site, but identifies a strong case of cooperativity where the ERG 6-mer CCGGAA is followed by the TEAD4 6-mer CATTCC site with an 8 bp gap. Logo of the PWM for this site is also shown. **b**, Comparison between CAP-SELEX PWMs and previously characterized specificities for the indicated TF pairs. This method has been used previously and its references are also indicated. CAP-SELEX models also shown in Fig. 1 are indicated by asterisks. Note that four out of five of the CAP-SELEX models are similar to the previously identified consensus sequences. The exception is ELK1–PAX5 consensus, that matches poorly both the CAP-SELEX motif and individual motifs for ELK1 and PAX5 (not shown). **c**, CAP-SELEX PWMs for TF pairs known to interact at protein level. Method used to identify the protein–protein interaction and its reference are also shown<sup>6,26–28,37,38,64–67</sup>.

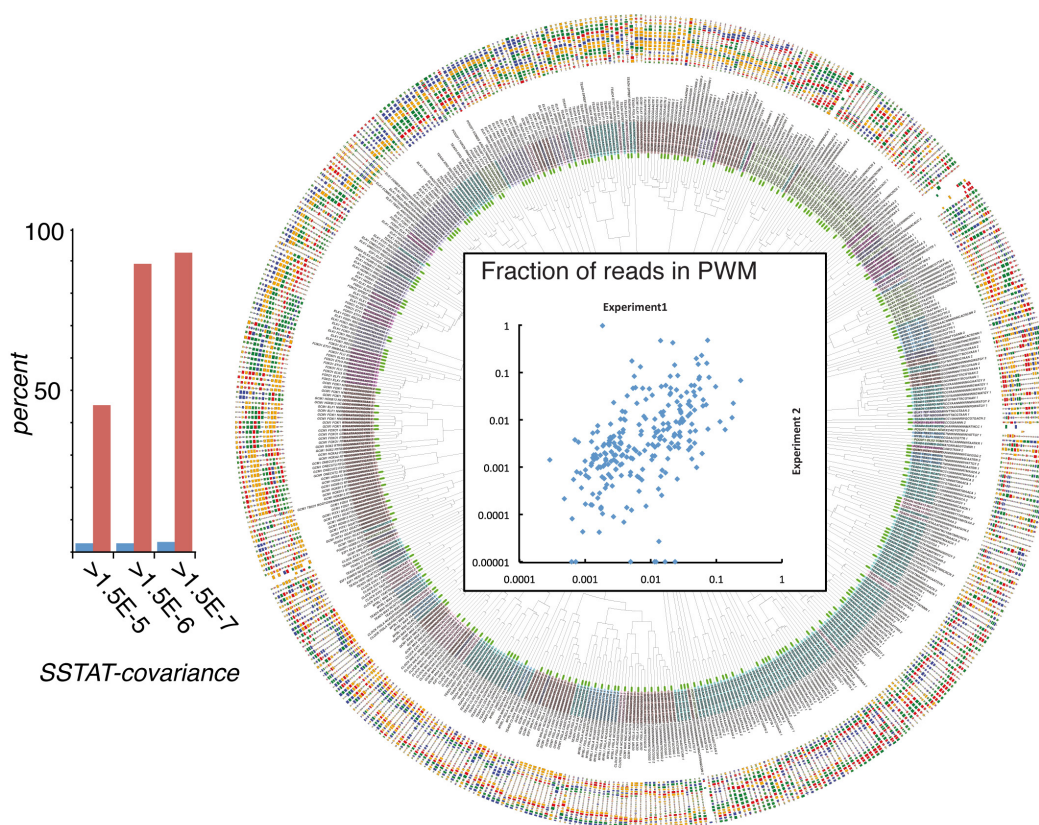


**Extended Data Figure 2 | Pairwise interaction matrix between TFs.** Columns indicate TF1 proteins, and rows TF2 proteins, subjected to the first and second affinity purifications, respectively. Pairs of TFs with a single spacing and orientation preference are indicated in dark green, and pairs with multiple preferred configurations in light green. White boxes indicate pairs that displayed weak or no interaction, and grey boxes cases where robust preference data was not recovered. Previously known interacting TF-pairs are indicated by a yellow outline (see Extended Data

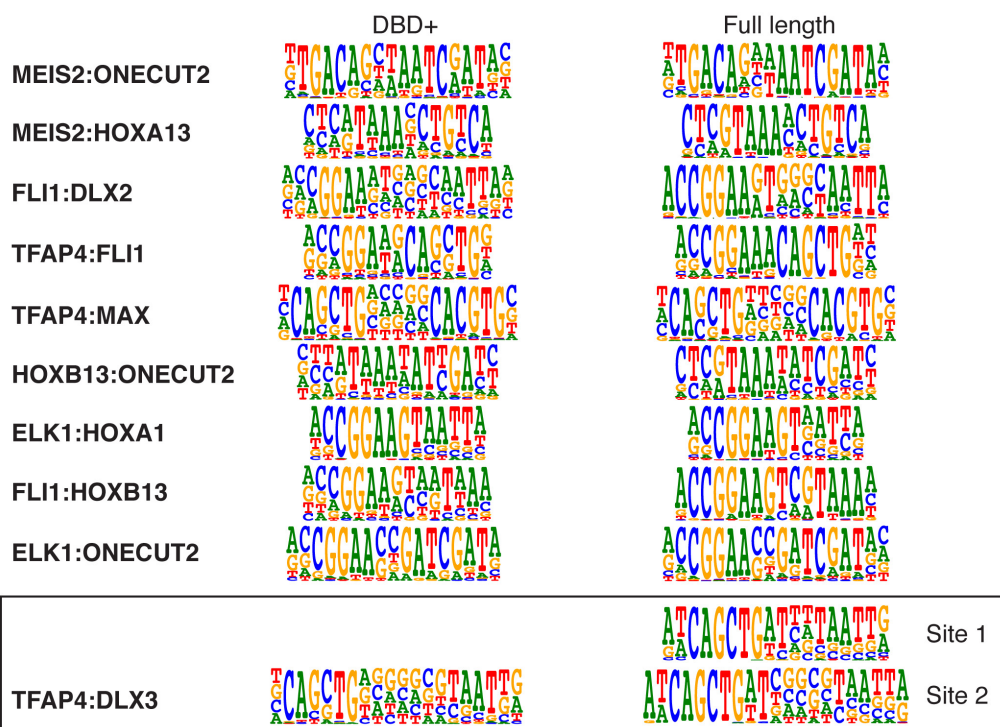
Fig. 1). Histograms show the counts for the interactions for each TF. Only TFs for which at least one clear interaction or independent binding was identified are included. The importance of including DNA in the interaction assay is highlighted by the fact that only four and five of the interactions detected are among those observed between 762 human or 877 mouse TF pairs identified using protein-protein interaction assays<sup>37</sup>, or compiled from literature<sup>36</sup>, respectively.



a

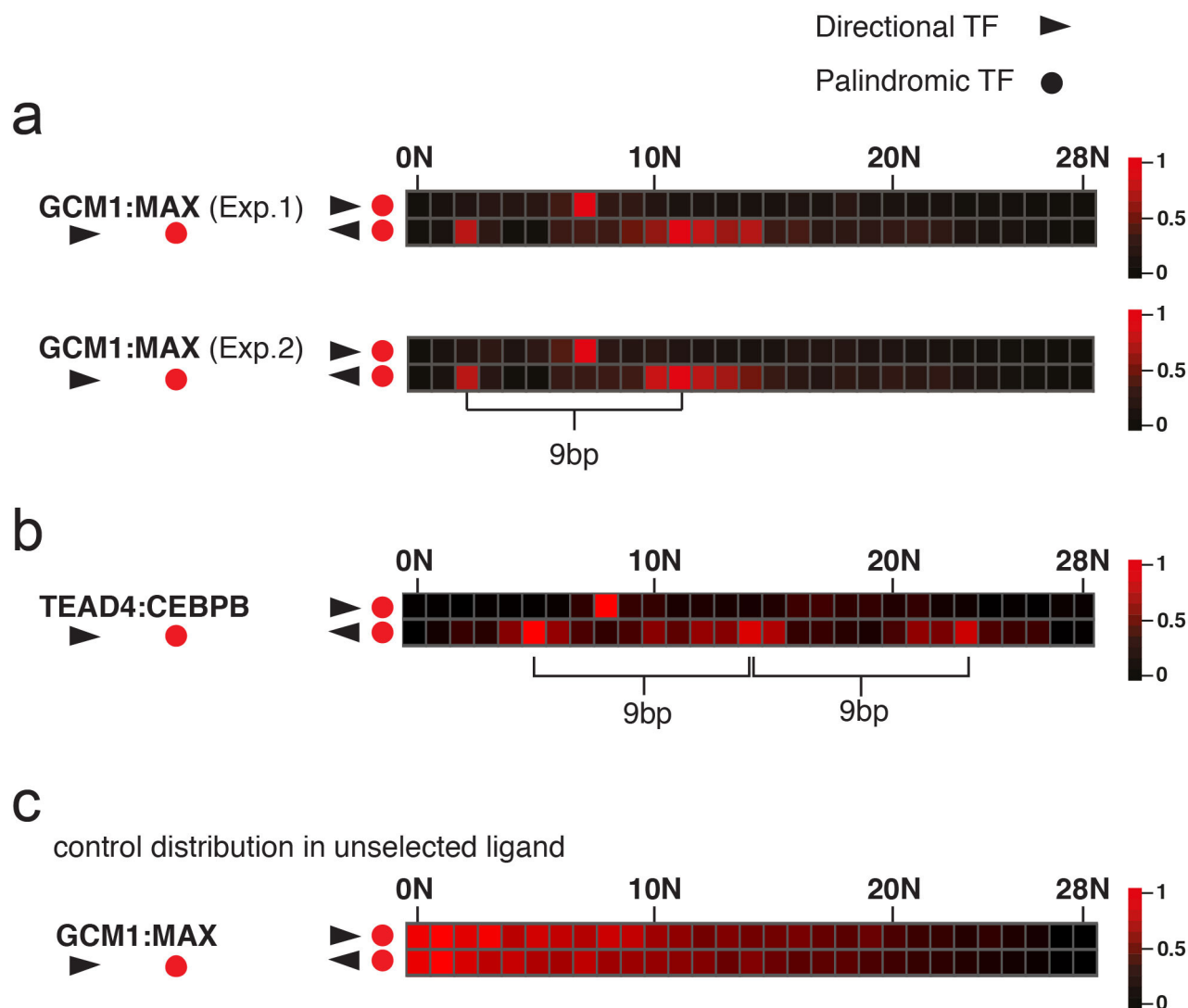


b



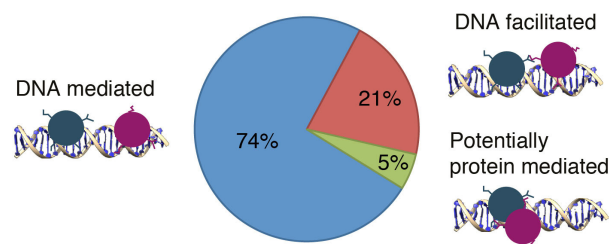
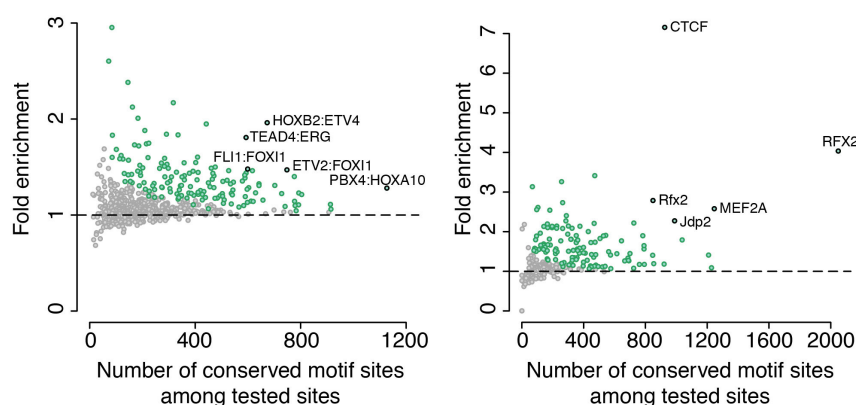
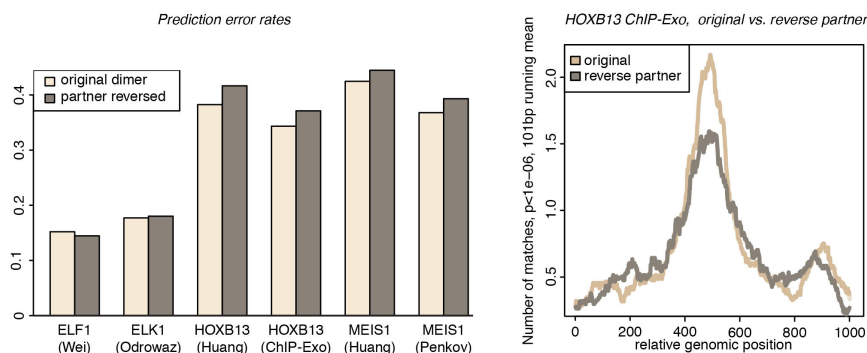
**Extended Data Figure 3 | CAP-SELEX reproducibility.** **a**, Replicate analysis of more than two hundred of the generated PWMs. The same seeds that had been used to generate PWMs for the primary experiments were used to seed new PWMs from the replicates. Left, red bars on the left show the percentage of the PWM pairs that are similar at the indicated cut-offs (measured as SSTAT covariance<sup>8,48</sup>). The highest threshold is the same used for identifying the dominating set of PWMs. Blue bars indicate fraction of all replicate PWMs that are similar using the same cut-off. Right, dendrogram

and barcode logos of all PWM pairs. Plot in the middle shows fraction of reads included in the same models in replicate 1 and 2. **b**, Validation of the CAP-SELEX analysis using shortened TF constructs (DBD+) by HT-SELEX using full-length protein mixtures (full length). Note that the same orientation and spacing is preferred in all but one of the cases. In one case (bottom), full-length proteins show the highest preference to a different spacing than that observed in CAP-SELEX; even in this case, the second-most preferred spacing is the one identified using CAP-SELEX.



**Extended Data Figure 4 | Long-range cooperativity.** Many experiments where TFs bound sites that were relatively far apart showed preferential binding to sites that are separated by approximately nine to ten bases. Heatmap (maximum count set to 1) representations showing frequency of occurrence of the representative 6-mers for TF pairs in all possible spacings (columns) and orientations (rows). **a**, Replicate experiment of GCM1 (black arrowhead) and MAX (red ball) pair show very similar preference for cooperatively bound representative 6-mers (see Supplementary Table 1). While one of the orientations shows preference for a single spacing, the second has two preferentially recognized regions

separated by ~9 bp. **b**, TEAD4–CEBPB pair shows a similar ~9 bp separation between three regions of preferred spacings (brackets). **c**, Very deep sequencing of the unselected input ligand does not show the same preference, instead counts decrease linearly as a function of gap length (due to decreasing number of available positions in the 40N random sequence). The mode of cooperativity seen in **a** and **b** appears similar to that reported by Kim *et al.*<sup>17</sup>. In addition to high-affinity sites, lower affinity spacings and orientations between TF pairs could be employed in fine-tuned transcriptional responses (see refs. 68, 69).

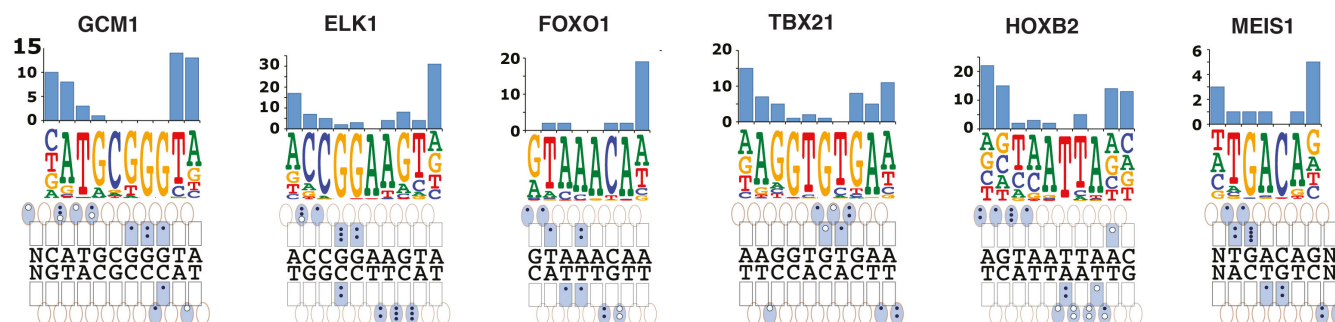
**a** Prevalence of interaction types based on structural analysis**b** Evolutionary conservation of the dimeric PWMs**c** Prediction of ChIP-seq peaks**Extended Data Figure 5 | CAP-SELEX motifs are conserved and enhance prediction of *in vivo* peaks.**

**a**, Pie chart showing the frequency of DNA-mediated, DNA-facilitated and potentially protein–protein interaction mediated heterodimers in the CAP-SELEX data set. Cooperativity between TFs can result from direct contacts between the proteins (protein-mediated), DNA-facilitated protein contacts (DNA-facilitated) or arise indirectly from DNA-mediated interactions<sup>17,34,39,40,70,71</sup>. The last type of cooperativity is caused by the DBD binding-induced changes in DNA shape, and do not involve other domains or direct contact between the proteins<sup>17,39,40</sup>. The dimers were classified to DNA-mediated, DNA-facilitated and potentially protein–protein interaction mediated classes manually, based on structural models shown in Supplementary Data Set 2. **b**, Conservation of the genomic sites recognized by the CAP-SELEX identified heterodimeric motifs (left) compared to monomeric and homodimeric sites identified by HT-SELEX (right, motifs from ref. 8). For each motif, ten thousand non-overlapping highest affinity sites within human constrained non-coding regions were selected and their conservation was tested. The fold enrichment (y axis), that is, the fraction of conserved sites among the motif sites

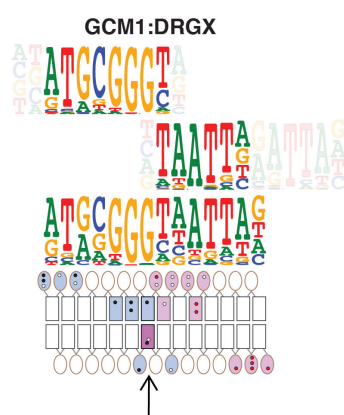
divided by the fraction of conserved sites among the control motif sites, is shown as a function of the number of conserved motif sites among the top ten thousand sites (x axis). The motifs that are significantly conserved (multiple testing adjusted  $P$  value  $< 0.05$ ) are marked green. Five motifs with lowest  $P$  values are also indicated. Note that ~50% of the HT-SELEX and ~25% of the CAP-SELEX motifs are conserved above the significance threshold. **c**, Inclusion of heterodimeric motifs improves prediction of ChIP-seq peaks. Left, the error rate of prediction of ChIP-seq peak positions using either the monomer motifs and CAP-SELEX dimers (light grey), or monomer motifs and control motifs where the partner of the indicated TF is reversed but not complemented (dark grey) are shown. Note that inclusion of the correct heterodimeric motifs decreases the prediction error rate in the cases of HOXB13 and MEIS1. The relatively modest effect is likely due to the fact that only a subset of heterodimers were identified in our study, and that ChIP-seq peak positions are also influenced by other factors such as nucleosome binding and chromatin structure. Right, number of PWM matches as a function of distance from HOXB13 ChIP-exo peaks. Note that using the original heterodimer motifs clearly outperforms the control motifs.



# a Summary of positions altered by heterodimer formation



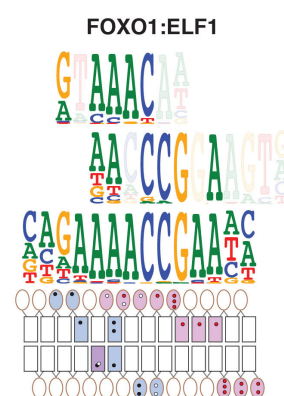
# b Contacts from both minor and major grooves



# c From homodimer to heterodimer



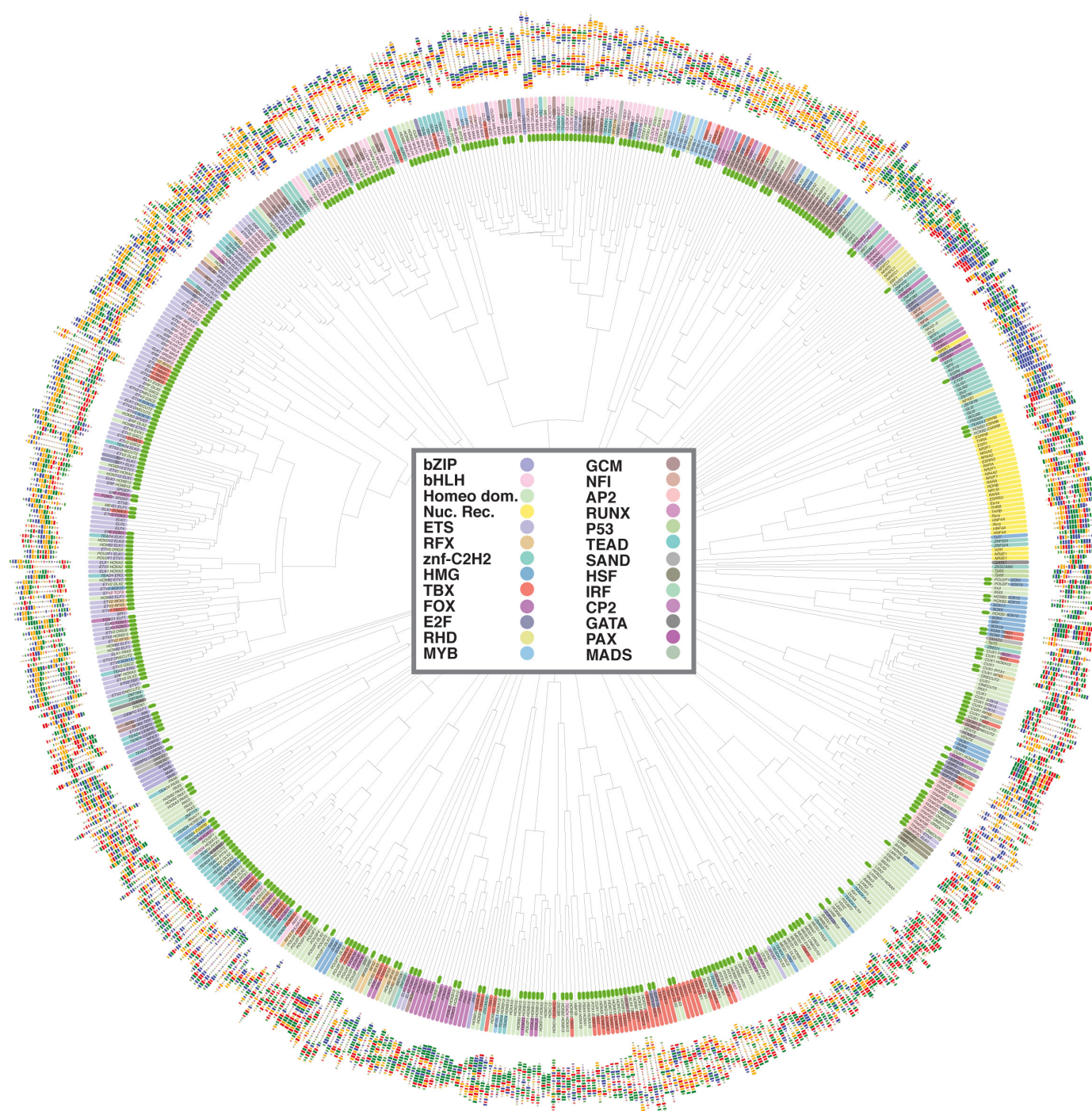
# d Binding positions cannot be assigned



**Extended Data Figure 6 | Heterodimers where the individual TF core recognition sites appear to overlap.** **a**, Composite site formation alters specificity at bases flanking the core TF site. TFs often directly read specific 'core' sequence motifs via hydrogen bonding to DNA bases. The sequences flanking this core are commonly read indirectly, through contacts to the sugar and phosphate backbone of DNA<sup>72–74</sup>. The backbone contacts specify a preferred DNA conformation, which then leads to a preference of a sequence that is optimal for stacking interactions between consecutive base pairs (reviewed in ref. 74). Figure shows summary of base positions whose specificity is affected in all composite sites identified in this study for the indicated TFs. Note that the bases comprising the core motif that is recognized by direct hydrogen bonds to the DNA bases are not commonly affected by heterodimer formation. In contrast, specificity at flanking positions that are recognized by contacts to the

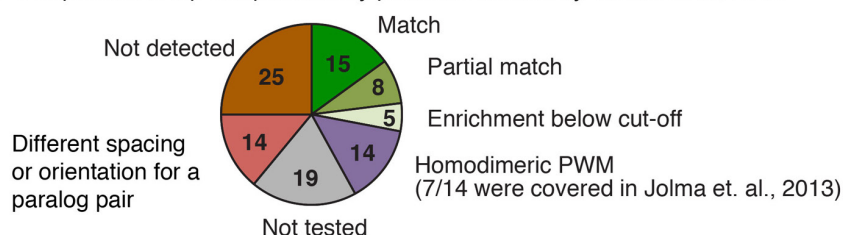
sugar or phosphate backbone of DNA are commonly altered by binding of the heterodimer partner. Hydrogen bonds contacts were determined based on the indicated (refs 29 and 30) or homologous TF structures (see Supplementary Table 3). **b**, A base (arrow) can be contacted both from the side of the major groove (black dot; G contacted by GCM1) and the minor groove (white dot; C contacted by DRGX homeodomain). **c**, A TF that can bind to a homodimeric site appears instead to bind as a heterodimer. A composite site is shown where HOXB2 appears to form a heterodimer with a monomer of RFX5. **d**, In some cases, the binding positions of the TFs cannot be unambiguously assigned based on the composite recognition sequence. In **a**, the annotation of hydrogen bond contacts is as described in main Fig. 2; in **b–d**, the major groove contacts of the left and right TFs are indicated in black and red dots, respectively.

*Representative PWMs of heterodimeric complexes and individual TFs*



**Extended Data Figure 7 | Specificities of individual TFs and heterodimer pairs.** Dendrogram shows motif similarities between the representative heterodimer and monomer motifs. Heterodimer models are indicated by green bars. Barcode logos for each factor are also shown. Centre of dendrogram shows the colour key for the TF families.



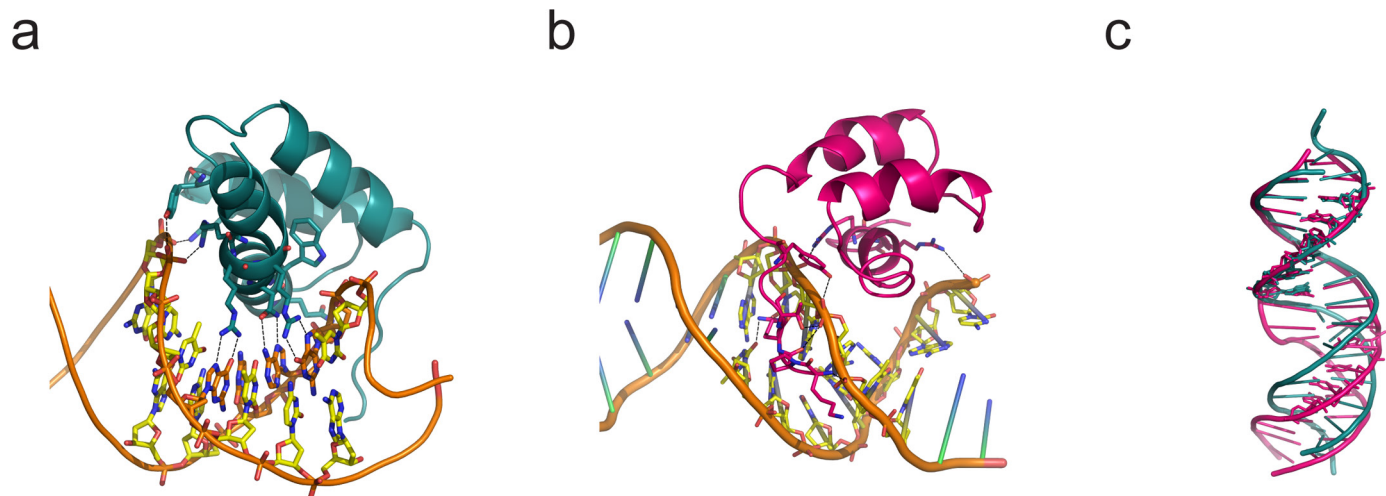
**a** Comparison of top computationally predicted models by Guturu *et al.*, 2013**b** Heterodimer SELEX result matches and partial matches to 100 most significant models

	Guturu <i>et al.</i> 2013 model	CAP-SELEX model	TF pair
FOX+{0}ETS+ _01			FOXO1:FLI1
PBX+{-1}HOX*9+ _01			PBX2:HOXA9
BHLH+{4}BARHL- _01			HOXB2:TCF3
BHLH+{5}BHLH+ _01			CLOCK:BHLHA15
BHLH+{0}NFAT- _01			ETV5:FIGLA
HOX+{1}ETS- _01			ETV2:DRGX
NANOG+{0}TBX+ _01			HOXA3:EOMES
BHLH+{4}BARHL- _02			ARNTL:PITX1
OCT-{1}SOX+ _01			POU2F1:SOX2
ETS+{3}HOX*9- _01			ELK1:HOXB13
ETS+{0}SOX+ _01			FLI1:DLX2
BHLH+{6}BHLH+ _01			TFAP4:MAX
HOX+{1}TBX- _01			HOXB2:TBX3
NANOG+{6}SOX- _01			HOXB2:SOX15
BHLH+{4}BHLH+ _01			ARNTL_BHLHA15
CUX-{-5}HOX*9+ _01			PBX4:HOXA10
ETS+{2}HOX+ _01			ELK1:HOXA3
ETS+{0}ETS- _01			FLI1:ETV7
BZIP-p+{1}ETS- _01			ETV2:TEF
NANOG+{4}NFAT+ _01			HOXB2:ETV7
SOX+{0}ETS- _01			ELK1:HOXA3
ETS+{4}HBP- _01			ELK1:HOXA3
HOX+{-3}GCM-s- _01			HOXB2:PITX1

**Extended Data Figure 8 | Comparison of CAP-SELEX models to models inferred from conserved genomic sequences.** **a**, Motifs that are very similar to the CAP-SELEX motifs are enriched and conserved. A previous study by Guturu *et al.*<sup>20</sup> made structural models for pairs of TFs to identify sterically possible configurations and predict sites that could be bound by such complexes. Enrichment of those target sites were then quantified in evolutionarily conserved noncoding regions over nonconserved control regions to infer putative target sites for cooperatively binding TFs. Pie chart shows comparison of top 100 most significant target sites predicted<sup>20</sup> to all heterodimeric PWMs generated in this study. 15 PWMs showed clear similarity to our heterodimeric PWMs (upper right, dark green slice),

8 were partially similar (green) and further 5 had enrichment for the site but under the threshold used in our study. We did not detect 25 motifs, and for 14 potential pairs, we identified a different spacing and orientation. This result is expected as we did not test all potential TF–TF pairs, and many TFs that bind to similar monomer sites prefer different dimer spacings and orientations. Finally, of the 100 Guturu *et al.*<sup>20</sup> top motifs, 33 were not analysed in our study (14 were homodimeric and no possible pair was tested for 19; for example, three of the pairs were predicted for pairs with a SMAD TF, and no SMAD TFs were tested in our study). **b**, Comparison of the 15 (top) and 8 (bottom, boxed) matching and partially matching PWMs, respectively.





**Extended Data Figure 9 | Detailed view of MEIS1 and MEIS1–DLX3 structures.** **a**, Contacts between MEIS1 (cyan) and DNA. **b**, Contacts between DLX3 (magenta) and DNA. **c**, Comparison of the DNA structures in MEIS1 homodimer (cyan) and MEIS1–DLX3 heterodimer (magenta). Note that the DNA bound to the heterodimer is more distorted.

# Histone H1 couples initiation and amplification of ubiquitin signalling after DNA damage

Tina Thorslund<sup>1\*</sup>, Anita Ripplinger<sup>1\*</sup>, Saskia Hoffmann<sup>1\*</sup>, Thomas Wild<sup>2\*</sup>, Michael Uckelmann<sup>3</sup>, Bine Villumsen<sup>1</sup>, Takeo Narita<sup>2</sup>, Titia K. Sixma<sup>3</sup>, Chunaram Choudhary<sup>2</sup>, Simon Bekker-Jensen<sup>1</sup> & Niels Møllgaard<sup>1</sup>

DNA double-strand breaks (DSBs) are highly cytotoxic DNA lesions that trigger non-proteolytic ubiquitylation of adjacent chromatin areas to generate binding sites for DNA repair factors. This depends on the sequential actions of the E3 ubiquitin ligases RNF8 and RNF168 (refs 1–6), and UBC13 (also known as UBE2N), an E2 ubiquitin-conjugating enzyme that specifically generates K63-linked ubiquitin chains<sup>7</sup>. Whereas RNF168 is known to catalyse ubiquitylation of H2A-type histones, leading to the recruitment of repair factors such as 53BP1 (refs 8–10), the critical substrates of RNF8 and K63-linked ubiquitylation remain elusive. Here we elucidate how RNF8 and UBC13 promote recruitment of RNF168 and downstream factors to DSB sites in human cells. We establish that UBC13-dependent K63-linked ubiquitylation at DSB sites is predominantly mediated by RNF8 but not RNF168, and that H1-type linker histones, but not core histones, represent major chromatin-associated targets of this modification. The RNF168 module (UDM1) recognizing RNF8-generated ubiquitylations<sup>11</sup> is a high-affinity reader of K63-ubiquitylated H1, mechanistically explaining the essential roles of RNF8 and UBC13 in recruiting RNF168 to DSBs. Consistently, reduced expression or chromatin association of linker histones impair accumulation of K63-linked ubiquitin conjugates and repair factors at DSB-flanking chromatin. These results identify histone H1 as a key target of RNF8–UBC13 in DSB signalling and expand the concept of the histone code<sup>12,13</sup> by showing that posttranslational modifications of linker histones can serve as important marks for recognition by factors involved in genome stability maintenance, and possibly beyond.

To explain mechanistically the critical role of UBC13 in the RNF8/RNF168 pathway, we generated human *UBC13* knockout cells using CRISPR–Cas9 technology<sup>14,15</sup> (Fig. 1a). As expected, loss of UBC13 abrogated the accumulation of RNF168, RNF168-dependent ubiquitin conjugates, and 53BP1 and other readers of these marks at DSB sites, while RNF8 recruitment was normal (Fig. 1b, c and Extended Data Fig. 1a, b). Reintroducing UBC13 into these cells restored 53BP1 focus formation, as did a fusion protein mimicking an RNF8–UBC13 E3–E2 complex<sup>16,17</sup>, independently of endogenous RNF8 (Fig. 1d and Extended Data Fig. 1c, d). In contrast, expression of RNF8 alone or other RNF8–E2 chimaeras failed to support 53BP1 accumulation at DSBs in these cells (Fig. 1d and Extended Data Fig. 1c). Unlike RNF8, overexpression of RNF168 was sufficient to restore 53BP1 recruitment to damaged DNA in *UBC13*-knockout cells (Extended Data Fig. 1e). These data suggest that UBC13 mainly cooperates with RNF8 but not RNF168 to catalyse K63 ubiquitylation of DSB-flanking chromatin to promote recruitment of repair factors.

To understand the molecular basis of this process, we used a tandem ubiquitin-binding entity ('K63-Super-UIM')<sup>18</sup> that showed remarkable specificity for binding to K63 linkages but not to other ubiquitin

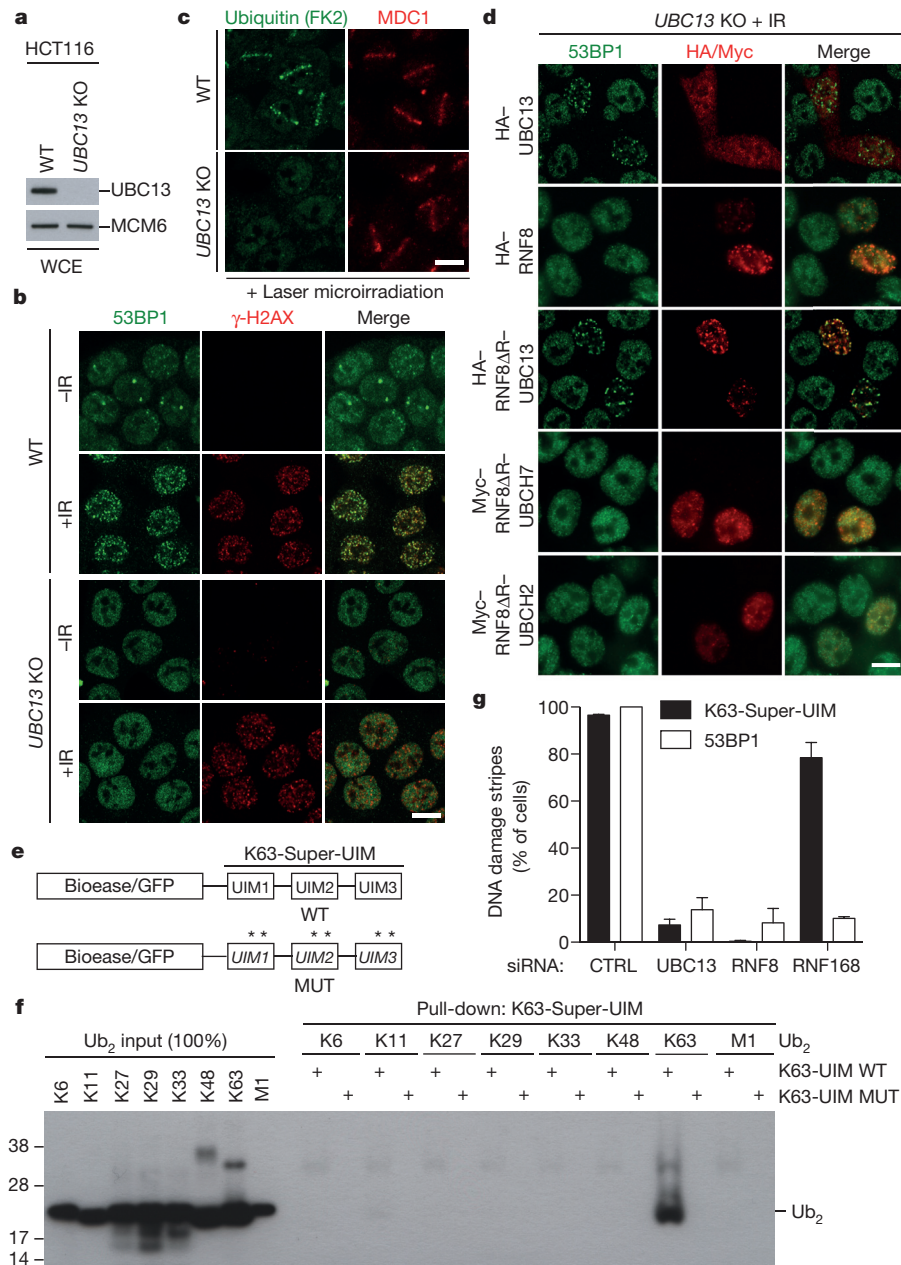
chain types (Fig. 1e, f). Indeed, green fluorescent protein (GFP)-tagged K63-Super-UIM, but not a ubiquitin-binding-deficient mutant (Fig. 1e, f), was efficiently recruited to microlaser- and ionizing radiation (IR)-generated DSBs (Extended Data Fig. 2a–c), thus providing an efficient sensor of DSB-associated K63 ubiquitylation<sup>19</sup>. Depletion of UBC13 or RNF8, but not RNF168, abrogated recruitment of the K63-Super-UIM, as well as an independently derived high-affinity binder of K63-linked ubiquitin chains, to DSB sites, and auto-ubiquitylation-generated K63 ubiquitin chains were abundantly present on RNF8 but not RNF168 (Fig. 1g and Extended Data Fig. 2b–f). This suggests that the roles of RNF8 and RNF168 in promoting DSB-associated chromatin ubiquitylation can be functionally uncoupled and that RNF8, but not RNF168, is a primary mediator of UBC13-dependent K63 ubiquitylation at DSB sites.

To identify the RNF8–UBC13-dependent ubiquitylation processes underlying RNF168 recruitment to DSBs, we first characterized the impact of UBC13 loss on global K63-linked ubiquitylation under steady-state conditions. While K63-linked ubiquitylation is thought to be mainly catalysed by UBC13 (ref. 7), the extent to which this E2 contributes to the total pool of K63 ubiquitin linkages in human cells is not known. To address this, we quantified the relative abundance of different ubiquitin chains in HCT116 wild-type and *UBC13*-knockout cells using stable isotope labelling by amino acids in cell culture (SILAC)-based mass spectrometry. Ablation of UBC13 decreased the global level of K63 ubiquitin linkages by ~50%, while the abundance of other chain types was not markedly affected (Fig. 2a). Using the di-glycine approach<sup>20,21</sup>, we quantified over 3,000 ubiquitylation sites, of which fewer than 1% showed a >2-fold decrease in *UBC13*-knockout cells (Extended Data Fig. 3a–d and Supplementary Table 1). Thus, UBC13 has little impact on the conjugation of the initial ubiquitin moiety to substrates, consistent with previous reports that UBC13 primarily acts at the ubiquitin chain elongation step<sup>22–24</sup>. To identify UBC13-dependent K63-ubiquitylated proteins, we analysed K63-Super-UIM pull-downs from wild-type and *UBC13*-knockout cells under stringent buffer conditions by mass spectrometry. We identified 371 proteins that showed >2-fold enrichment in wild-type cells, several of which are known targets of K63 ubiquitylation (Extended Data Figs 3e–h, 4a, b and Supplementary Table 2).

We extended this proteomic strategy to search for the chromatin-bound substrate(s), whose K63 ubiquitylation by RNF8–UBC13 promotes recruitment of RNF168 and downstream factors to DSB sites (Extended Data Fig. 4c). Surprisingly, only few cellular proteins reproducibly displayed elevated K63-linked ubiquitylation upon IR-induced DSBs, and H1 linker histones, but not core histones, were major chromatin-associated factors showing such behaviour (Extended Data Fig. 4d). Using the K63-Super-UIM we confirmed biochemically for two endogenous histone H1 isoforms (H1.2 and H1x) that they were modified with K63-linked ubiquitin chains and, more importantly,

<sup>1</sup>Ubiquitin Signaling Group, Protein Signaling Program, The Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Blegdamsvej 3B, DK-2200 Copenhagen, Denmark. <sup>2</sup>Proteomics Program, The Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Blegdamsvej 3B, DK-2200 Copenhagen, Denmark. <sup>3</sup>Division of Biochemistry, Cancer Genomics Center, Netherlands Cancer Institute, 1066 CX Amsterdam, the Netherlands.

\*These authors contributed equally to this work.



**Figure 1 | RNF8 but not RNF168 is a primary mediator of UBC13-dependent protein recruitment to DSB-flanking chromatin.** **a**, Immunoblot analysis of HCT116 wild-type (WT) and *UBC13*-knockout (KO) cells. MCM6 was a loading control. WCE, whole-cell extract. **b–d**, Representative images of HCT116 wild-type and *UBC13*-knockout cells exposed to IR or laser microirradiation ( $n = 3$  experiments). Cells in **d** were transfected with the indicated expression constructs (Extended Data Fig. 1c). HA, haemagglutinin. **e, f**, Binding of recombinant wild-type and mutant (MUT) K63-Super-UIM to

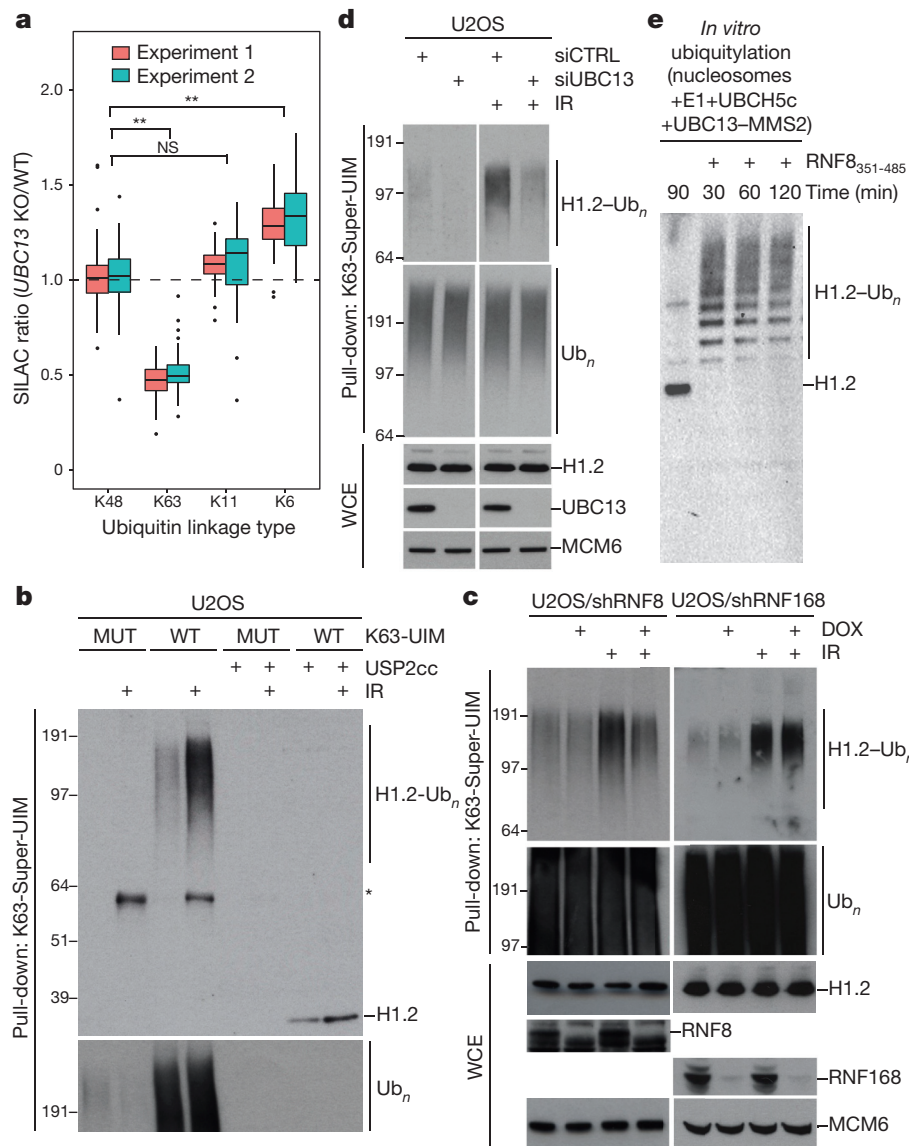
di-ubiquitin (Ub<sub>2</sub>) linkages. The migration of molecular weight markers (kDa) is indicated on the left. **g**, Quantification of GFP-K63-Super-UIM and 53BP1 accumulation at DSBs in U2OS/GFP-K63-Super-UIM cells transfected with the indicated siRNAs. CTRL, control. Data are mean  $\pm$  standard deviation (s.d.) from three independent experiments. Representative images are shown in Extended Data Fig. 2b. Scale bars, 10  $\mu$ m. **a, f**, Uncropped blots are shown in Supplementary Fig. 1.

that these ubiquitylations were markedly upregulated after DSBs (Fig. 2b and Extended Data Fig. 5a, b). Basal levels of H1 K63 poly-ubiquitylation in non-irradiated cells dropped substantially upon serum starvation (Extended Data Fig. 5a), suggesting that they result mostly from endogenous DNA damage generated by cell-cycle-dependent processes. Little if any K63-linked ubiquitylation of core histones was detectable, regardless of whether DSBs had been inflicted or not (Extended Data Fig. 5c), thus H1 appears unique among histones in undergoing robust DSB-stimulated K63 ubiquitylation. Importantly, the DNA-damage-induced increase in K63-linked poly-ubiquitylation of H1 was dependent on RNF8 and UBC13 but not RNF168 (Fig. 2c, d), whereas none of these enzymes appreciably affect

ed H1 monoubiquitylation (Extended Data Fig. 5d). Moreover, mass spectrometry analysis showed that H1 was a major factor co-purifying with endogenous RNF8 (Extended Data Fig. 5e). Finally, recombinant RNF8 catalysed robust ubiquitylation of nucleosome-associated H1 with UBC13 *in vitro* (Fig. 2e). These data suggest that histone H1 is a major chromatin-associated substrate of RNF8–UBC13.

To address whether linker histones represent critical RNF8 substrates in DSB signalling, we downregulated overall H1 levels using a short interfering RNA (siRNA) cocktail targeting different isoforms (Extended Data Fig. 6a, b). Under these conditions, accumulation of K63-linked ubiquitin chains, RNF168, 53BP1 and BRCA1, but not MDC1, at DSB sites was clearly diminished (Fig. 3a–d and Extended





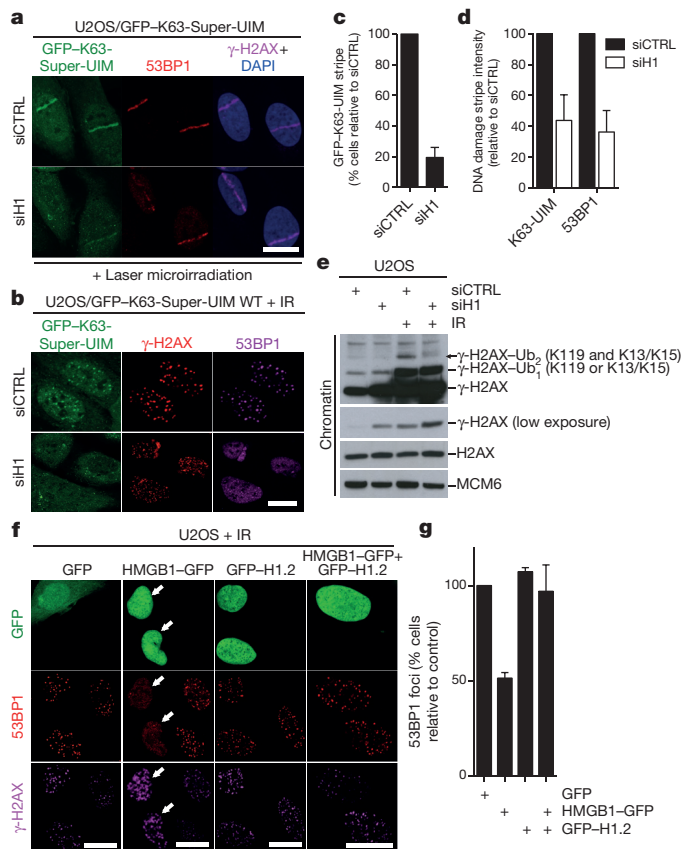
**Figure 2 | H1-type linker histones are major chromatin-bound targets of RNF8-UBC13-dependent K63 ubiquitylation.** **a**, Tukey boxplot showing levels of ubiquitin linkages in HCT116 *UBC13*-knockout (KO) cells relative to wild-type (WT) cells, quantified by SILAC-based proteomics. Data from two independent experiments are shown. \*\* $P < 0.01$ ; NS, not significant (Welch's  $t$ -test). **b–d**, Pull-down analysis of IR-induced K63-linked ubiquitylation of histone H1.2 in U2OS cells (**b**, **d**) or derivative cell lines expressing RNF8 or RNF168 short hairpin RNAs (shRNAs) in a doxycycline (DOX)-inducible

manner (**c**) using recombinant wild-type or mutant (MUT) K63-Super-UIM. Where indicated, ubiquitin conjugates were digested with recombinant USP2 catalytic domain (USP2cc) before SDS–polyacrylamide gel electrophoresis (SDS–PAGE). Asterisk indicates a non-specific band. **e**, *In vitro* ubiquitylation of purified H1-containing oligonucleosomes by RNF8<sub>351–485</sub> and indicated factors. WCE, whole-cell extract. The migration of molecular weight markers (kDa) is indicated on the left. **b–e**, Uncropped blots are shown in Supplementary Fig. 1.

Data Fig. 6c–g), suggesting that loss of H1 uncouples ubiquitin-dependent DSB signalling at the level of RNF168 accrual. Consistently, like RNF8 or RNF168 knockdown, H1 downregulation suppressed DSB-induced H2A ubiquitylation catalysed by RNF168 (Fig. 3e and Extended Data Fig. 6h)<sup>8,10</sup>. As an alternative strategy to interfere with H1 functionality, we overexpressed the high-mobility group protein HMGB1, which competes with H1 for chromatin binding in a non-site-specific manner<sup>25</sup>. Elevated HMGB1 levels impaired K63 ubiquitylation and 53BP1 accumulation at DSB sites despite enhancing  $\gamma$ -H2AX formation, and this could be rescued by co-overexpression of H1 or by wild-type but not catalytically inactive RNF168 (Fig. 3f, g and Extended Data Fig. 7a–c). A similar strong increase in  $\gamma$ -H2AX formation accompanied by only mildly elevated 53BP1 accumulation at DSBs, indicating a partial uncoupling of these events, was reported for murine cells lacking three of six H1 isoforms<sup>26</sup>. Together, these data suggest that H1-type histones are functionally critical targets

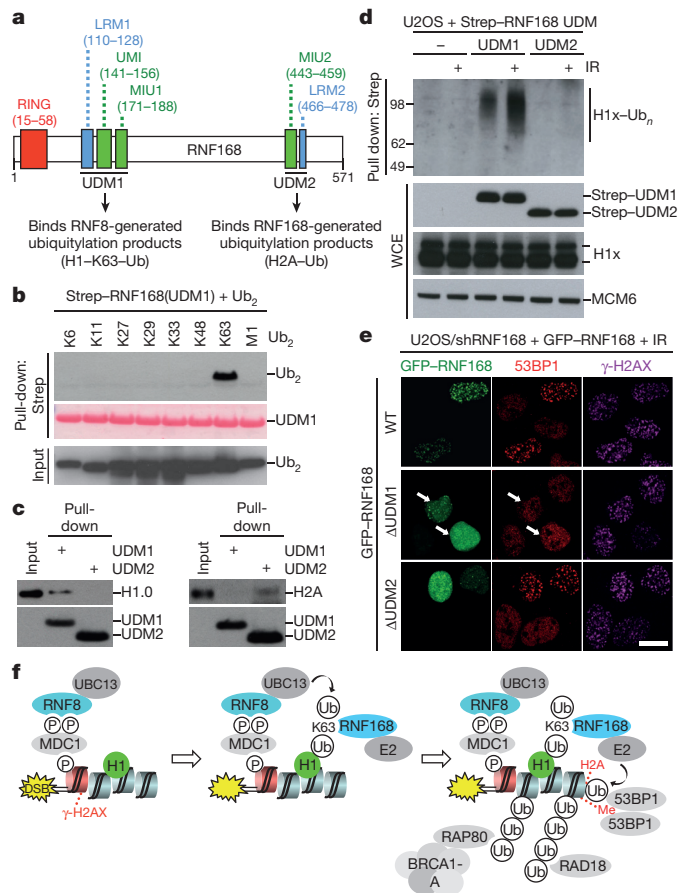
of RNF8- and UBC13-dependent K63 ubiquitylation at DSB-flanking chromatin areas.

We asked whether K63-ubiquitylated H1 provides a platform for initial, RNF8-UBC13-dependent RNF168 recruitment to DSBs. RNF168 contains two DSB-targeting modules, both of which contain ubiquitin-binding domains<sup>4,5,27</sup> (Fig. 4a). While the carboxy-terminal LRM2-MIU2 region, which we termed ubiquitin-dependent DSB recruitment module 2 (UDM2), recognizes RNF168-catalysed forms of ubiquitylated H2A, thereby effectively enabling RNF168 to autonomously propagate H2A ubiquitylation at DSB sites once recruited, its LRM1-UMI-MIU1 region (UDM1) has been suggested to bind as yet undefined RNF8-generated ubiquitylation products required for initial relocalization of RNF168 to break sites (Fig. 4a)<sup>11</sup>. Therefore, if K63-ubiquitylated H1 functions as a recruitment platform for RNF168 at DSB-modified chromatin, the UDM1 module should be capable of recognizing this modified form of H1. Consistent with this idea,



**Figure 3 | Histone H1 is required for K63-linked ubiquitylation and RNF168-dependent protein retention at DSB sites.** **a, b,** Representative images of GFP-K63-Super-UIM-expressing cells transfected with siRNAs. CTRL, control; DAPI, 4',6-diamidino-2-phenylindole. **c,** Proportion of cells with visible accumulation of GFP-K63-Super-UIM at laser-induced DSBs (**a**). **d,** Intensity of GFP-K63-Super-UIM or 53BP1 accumulation at laser-induced DSBs normalized to  $\gamma$ -H2AX signal (**a**). **e,** Analysis of IR-induced  $\gamma$ -H2AX ubiquitylation (Ub) by RNF168 (marked by arrow) in chromatin fractions of U2OS cells transfected with the indicated siRNAs. **f,** Representative images of U2OS cells expressing the indicated constructs and exposed to IR. HMGB1 overexpression impairs 53BP1 recruitment to DSBs (marked by arrows). **g,** Quantification of data in **f**. Data in **c** and **d** are mean  $\pm$  s.d. from three independent experiments. Data in **g** are mean  $\pm$  standard error of the mean (s.e.m.) from two independent experiments. Scale bars, 10  $\mu$ m. **e,** Uncropped blots are shown in Supplementary Fig. 1.

we found that UDM1 bound to K63-linked ubiquitin but no other chain types (Fig. 4b). RNF168 UDM2 also interacted with K63 linkages, albeit with much lower affinity, and unlike UDM1 it also recognized other ubiquitin chains (Extended Data Fig. 8a, b). This might reflect the fact that UDM1, but not UDM2, has two adjacent ubiquitin-binding domains (UMI and MIU1), which together may confer specific binding to K63-linked chains. We also found that UDM1 interacted with histone H1 but not H2A *in vitro*, whereas UDM2 displayed the inverse preference, binding only to H2A, as previously shown<sup>11</sup> (Fig. 4c). The LRM1 part of UDM1 has been hypothesized to impart target specificity to the ubiquitin-binding affinity of this module<sup>11</sup>. Consistently, the H1-binding ability of UDM1 was at least partially mediated by the LRM1 motif (Fig. 4a), which interacted strongly with H1 *in vitro*, unlike LRM2 (Extended Data Fig. 8c). However, additional sequence elements within UDM1 were also able to bind H1 (data not shown), possibly owing to the very acidic nature of the UDM1 region (Extended Data Fig. 8d), which could facilitate robust interaction with the highly basic H1 proteins<sup>28</sup>. Unlike UDM1, UDM2 interacted with neither modified nor unmodified forms of H1 (Fig. 4c, d). When expressed in cells, the UDM1 domain efficiently bound to



**Figure 4 | The RNF168 UDM1 domain is a high-affinity reader of K63-ubiquitylated H1.** **a,** Composition and reported functions of ubiquitin-dependent DSB recruitment modules (UDMs) in human RNF168. **b,** Binding of recombinant UDM1 to di-ubiquitin (Ub<sub>2</sub>) linkages. **c,** Binding of recombinant UDM1 and UDM2 to purified histones (H1.0 and H2A). **d,** Pull-down assays of Strep-tagged UDM1 and UDM2 modules expressed in U2OS cells. The migration of molecular weight markers (kDa) is indicated on the left. WCE, whole-cell extract. **e,** Representative images of shRNF168-expressing cells transfected with GFP-RNF168 constructs and exposed to IR ( $n = 2$  experiments). Deletion of UDM1 ( $\Delta$ UDM1) impairs restoration of 53BP1 foci by GFP-RNF168 (indicated by arrows). Scale bar, 10  $\mu$ m. **f,** Model of RNF8-UBC13 function in ubiquitin-dependent signalling after DSBs. Me, methyl group; P, phosphate. **b–d,** Uncropped blots are shown in Supplementary Fig. 1.

high-molecular-weight forms of endogenous H1 isoforms in an IR-stimulated manner, suggesting that they correspond to K63-ubiquitylated H1 (Fig. 4d and Extended Data Fig. 8e). Supporting this, point mutations in the UMI or MIU1 domains that impair their ubiquitin-binding activity and RNF168 recruitment to DSBs<sup>11,27</sup> markedly reduced the binding of UDM1 to modified H1 (Extended Data Fig. 8f). Moreover, UDM1 did not interact with ubiquitylated forms of core histones, in sharp contrast to its binding to ubiquitylated H1 (Extended Data Fig. 9a). Finally, in-frame deletion of UDM1 but not UDM2 impaired the ability of ectopically expressed RNF168 to promote IR-induced 53BP1 foci in cells lacking endogenous RNF168 (Fig. 4e). We conclude that the RNF168 UDM1 is a high-affinity recognition module for K63-ubiquitylated histone H1. While the isolated UDM1 domain was distributed pan-cellularly and interacted with a range of proteins when overexpressed in cells, the only such protein showing DSB-stimulated K63 polyubiquitylation was an H1 isoform (Extended Data Fig. 9b, c), further suggesting that ubiquitylated H1 is a major receptor for RNF168 at DSB-modified chromatin.

Collectively, our findings suggest an integrated model for how RNF8 and UBC13 promote ubiquitin-dependent protein recruitment

to DSB sites. We propose that H1-type linker histones represent key chromatin-associated RNF8 substrates whose UBC13-dependent K63-linked ubiquitylation at DSB-containing chromatin provide an initial binding platform for RNF168 via the UDM1 (Fig. 4f). RNF168 then ubiquitylates H2A at K13/K15 and possibly other proteins to trigger recruitment of DSB repair factors. The notable absence of K63-linked ubiquitin chains on H2A suggests that RNF168 does not efficiently modify H2A in conjunction with UBC13, but may instead catalyse formation of other ubiquitin chains or monoubiquitylation of the H2A K13/K15 mark. Indeed, RNF168 was recently shown to catalyse the formation of K27-linked ubiquitin chains, at least when over-expressed<sup>29</sup>. It is possible, however, that RNF168 cooperates with UBC13 in K63-linked ubiquitylation of other factors at damaged chromatin. While we observed no marked DSB-induced change in overall nuclear H1 mobility and no overt enrichment or depletion of H1 isoforms at DSB sites, the K63-ubiquitylated forms of H1 proteins were more loosely associated with chromatin than their unmodified counterparts (Extended Data Fig. 10a, b; data not shown). This suggests that in addition to triggering RNF168 recruitment, the DSB-associated K63 ubiquitylation of H1 may play a part in facilitating chromatin remodelling to allow efficient repair.

H1-type histones consist of almost 30% lysine residues<sup>28</sup>, and proteomic studies have shown that many of these can be ubiquitylated<sup>20,21,30</sup>. Because UBC13 only generates few *de novo* ubiquitylation marks, it is conceivable that RNF8–UBC13 mainly extends pre-existing H1 ubiquitylations in response to DNA damage, rendering identification of the key H1 sites targeted by RNF8–UBC13-dependent K63 ubiquitylation challenging. Although numerous post-translational modifications (PTMs) have been mapped on different H1 isoforms<sup>28</sup>, these marks have not so far been connected directly to the ‘histone code’ specifying protein recruitment to different chromatin states<sup>12,13</sup>. In addition to providing the first insights into the functional role of H1 ubiquitylation, our findings also show that linker histones can indeed form part of a dynamic histone code for DNA repair, wherein RNF8 and RNF168 function as writer and reader, respectively, of K63-ubiquitylated H1. The multitude of site-specific H1 PTMs raises the possibility that some of these marks may have an integral role in the histone code, an area of study warranting further investigation.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 3 December 2014; accepted 20 August 2015.**

**Published online 21 October 2015.**

- Mailand, N. *et al.* RNF8 ubiquitylates histones at DNA double-strand breaks and promotes assembly of repair proteins. *Cell* **131**, 887–900 (2007).
- Huen, M. S. *et al.* RNF8 transduces the DNA-damage signal via histone ubiquitylation and checkpoint protein assembly. *Cell* **131**, 901–914 (2007).
- Kolas, N. K. *et al.* Orchestration of the DNA-damage response by the RNF8 ubiquitin ligase. *Science* **318**, 1637–1640 (2007).
- Doil, C. *et al.* RNF168 binds and amplifies ubiquitin conjugates on damaged chromosomes to allow accumulation of repair proteins. *Cell* **136**, 435–446 (2009).
- Stewart, G. S. *et al.* The RIDDLE syndrome protein mediates a ubiquitin-dependent signaling cascade at sites of DNA damage. *Cell* **136**, 420–434 (2009).
- Jackson, S. P. & Durocher, D. Regulation of DNA damage responses by ubiquitin and SUMO. *Mol. Cell* **49**, 795–807 (2013).
- Hofmann, R. M. & Pickart, C. M. Noncanonical MMS2-encoded ubiquitin-conjugating enzyme functions in assembly of novel polyubiquitin chains for DNA repair. *Cell* **96**, 645–653 (1999).
- Mattioli, F. *et al.* RNF168 ubiquitinates K13-15 on H2A/H2AX to drive DNA damage signaling. *Cell* **150**, 1182–1195 (2012).
- Fradet-Turcotte, A. *et al.* 53BP1 is a reader of the DNA-damage-induced H2A Lys 15 ubiquitin mark. *Nature* **499**, 50–54 (2013).

- Gatti, M. *et al.* A novel ubiquitin mark at the N-terminal tail of histone H2As targeted by RNF168 ubiquitin ligase. *Cell Cycle* **11**, 2538–2544 (2012).
- Panier, S. *et al.* Tandem protein interaction modules organize the ubiquitin-dependent response to DNA double-strand breaks. *Mol. Cell* **47**, 383–395 (2012).
- Kouzarides, T. Chromatin modifications and their function. *Cell* **128**, 693–705 (2007).
- Jenuwein, T. & Allis, C. D. Translating the histone code. *Science* **293**, 1074–1080 (2001).
- Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
- Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).
- Bekker-Jensen, S. *et al.* HERC2 coordinates ubiquitin-dependent assembly of DNA repair factors on damaged chromosomes. *Nature Cell Biol.* **12**, 80–86, 1–12 (2010).
- Huen, M. S. *et al.* Noncanonical E2 variant-independent function of UBC13 in promoting checkpoint protein assembly. *Mol. Cell Biol.* **28**, 6104–6112 (2008).
- Sims, J. J. *et al.* Polyubiquitin-sensor proteins reveal localization and linkage-type dependence of cellular ubiquitin signaling. *Nature Methods* **9**, 303–309 (2012).
- van Wijk, S. J. *et al.* Fluorescence-based sensors to monitor localization and functions of linear and K63-linked ubiquitin chains in cells. *Mol. Cell* **47**, 797–809 (2012).
- Wagner, S. A. *et al.* A proteome-wide, quantitative survey of *in vivo* ubiquitylation sites reveals widespread regulatory roles. *Mol. Cell. Proteomics* **10**, M111.013284 (2011).
- Kim, W. *et al.* Systematic and quantitative assessment of the ubiquitin-modified proteome. *Mol. Cell* **44**, 325–340 (2011).
- Petroski, M. D. *et al.* Substrate modification with lysine 63-linked ubiquitin chains through the UBC13-UEV1A ubiquitin-conjugating enzyme. *J. Biol. Chem.* **282**, 29936–29945 (2007).
- Christensen, D. E., Brzovic, P. S. & Kleit, R. E. E2-BRCA1 RING interactions dictate synthesis of mono- or specific polyubiquitin chain linkages. *Nature Struct. Mol. Biol.* **14**, 941–948 (2007).
- Windheim, M., Pegg, M. & Cohen, P. Two different classes of E2 ubiquitin-conjugating enzymes are required for the mono-ubiquitination of proteins and elongation by polyubiquitin chains with a specific topology. *Biochem. J.* **409**, 723–729 (2008).
- Catez, F., Ueda, T. & Bustin, M. Determinants of histone H1 mobility and chromatin binding in living cells. *Nature Struct. Mol. Biol.* **13**, 305–310 (2006).
- Murga, M. *et al.* Global chromatin compaction limits the strength of the DNA damage response. *J. Cell Biol.* **178**, 1101–1108 (2007).
- Pinato, S., Gatti, M., Scanduzzi, C., Confalonieri, S. & Penengo, L. UMI, a novel RNF168 ubiquitin binding domain involved in the DNA damage signaling pathway. *Mol. Cell Biol.* **31**, 118–126 (2011).
- Harshman, S. W., Young, N. L., Parthun, M. R. & Freitas, M. A. H1 histones: current perspectives and challenges. *Nucleic Acids Res.* **41**, 9593–9609 (2013).
- Gatti, M. *et al.* RNF168 promotes noncanonical K27 ubiquitination to signal DNA damage. *Cell Reports* **10**, 226–238 (2015).
- Povlsen, L. K. *et al.* Systems-wide analysis of ubiquitylation dynamics reveals a key role for PAF15 ubiquitylation in DNA-damage bypass. *Nature Cell Biol.* **14**, 1089–1098 (2012).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank D. Durocher and M. Bianchi for providing reagents and J. Lukas for helpful discussions. This work was supported by grants from the Novo Nordisk Foundation (grants NNF14CC0001 and NNF12OC0002114), European Research Council, Nederlandse Organisatie voor Wetenschappelijk Onderzoek-Chemische Wetenschappen (NWO-CW), The Danish Cancer Society, and The Danish Council for Independent Research.

**Author Contributions** T.T. initiated the project, designed and performed cell biological and biochemical experiments, and analysed data; A.R. and S.H. performed cell biological and biochemical experiments and analysed data; T.W. generated UBC13-knockout cells, designed and performed mass spectrometry experiments and analysed the data; M.U. performed *in vitro* ubiquitylation assays with purified nucleosomes; B.V. helped T.T. with biochemical experiments; T.N. analysed mass spectrometry data; T.K.S. supervised M.U.; C.C. supervised T.W. and T.N., designed mass spectrometry experiments and analysed the data; S.B.-J. performed and designed experiments and analysed data; N.M. conceived and supervised the project, designed experiments, analysed data and wrote the manuscript with input from the other authors.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to N.M. (niels.mailand@cpr.ku.dk).



## METHODS

**Plasmids.** cDNAs encoding K63-Super-UIM (wild type and mutant) and the Vps27-based K63 binder<sup>18</sup> containing C-terminal His<sub>6</sub> tags were produced as synthetic genes (Eurofins) and inserted into pDONR221 by BP reactions (Invitrogen). By means of LR reactions (Invitrogen) the inserts were then transferred to the Champion pET104 BioEase Gateway Biotinylation System (Invitrogen) for recombinant protein production or pcDNA-DEST53 (Invitrogen) for GFP-tagged constitutive mammalian expression. For inducible expression of GFP-tagged K63-Super-UIM, the GFP-K63-Super-UIM complementary DNA was inserted into pcDNA4/TO (Invitrogen). Plasmids encoding HA-tagged wild-type and catalytically inactive (CI) RNF8 (C403S), wild-type and CI (C16S/C19S) forms of RNF168, and UBC13, as well as chimaeras between RNF8 and different E2 enzymes were described previously<sup>1,4,16</sup>. The 'FHA mutation (R42A) in HA-RNF8AR-UBC13 was generated by site-directed mutagenesis. RNF8 constructs were made resistant to RNF8-siRNA by introducing three silent mutations (bold) in the siRNA targeting sequence (5'-TGCGGAGTA TGAGTACGAG-3') in the plasmids by site-directed mutagenesis. The RNF168 UDM1 (amino acids 110–201) and UDM2 (amino acids 419–487) fragments were amplified by PCR and inserted into either pTriEx-5 (Novagen) for Strep- and His-tagged expression in *Escherichia coli* and mammalian cells, or pEGFP-C1 (Clontech) for expression of GFP-tagged versions. The Strep-RNF168 UDM1 mutants used in this study (\*UMI (L149A) and \*MIU1 (A179G)) were generated using the QuikChange site-directed mutagenesis kit (Stratagene). Constructs encoding GFP-H1 isoforms were cloned by inserting the respective cDNAs into the BglII and BamHI sites of pEGFP-C1 (Clontech). A plasmid encoding HMGB1-GFP was provided by M. Bianchi. A Flag-HMGB1 expression construct was generated by inserting the HMGB1 open reading frame (ORF) into pFlag-CMV2 (Sigma). All constructs were verified by sequencing. Plasmid transfections were done with FuGene 6 (Promega) or Genejuice (Novagene), siRNA transfections were done with Lipofectamine RNAiMAX (Invitrogen), according to the manufacturers' instructions.

**siRNA.** siRNA sequences used in this study were as follows. Non-targeting control (CTRL), 5'-GGGAUACCUAGACGUUCUATT-3'; UBC13, 5'-GAGCAU GGACUAGGCUAUAATT-3'; RNF8, 5'-UGCGGAGUAUGAAUUGAATT-3'; RNF168, 5'-GUGGAACUGUGGACGAUAATT-3' or 5'-GGCGAAGAGCG AUGGAAGATT-3'; histone H1(#1), 5'-GCUACGACGUGGAGAAGAATT-3'; H1(#2), 5'-GCUCCUUUAAACUCAACAATT-3'; H1(#3), 5'-GAAGCC AAGCCCAAGGUUATT-3'; H1(#4), 5'-CCUUAUAAACUCAACAAGAATT-3'; H1(#5), 5'-CCUUAUAAACUCAACAAGAATT-3'; H1(#6), 5'-UCAAGAG CCUGGUGAGCAATT-3'; H1(#7), 5'-GGACCAAGAAAGUGGCCAATT-3'; H1(#8), 5'-GCAUCAAGCUGGGUCUCAATT-3'; H1(#9), 5'-CAGUGAAA CCCAAAGCAAATT-3'; H1(#10) (specific for H1x), 5'-CCUUAAGCUC AACCCGAATT-3'; 53BP1, 5'-GAACGAGGAGACGGUAAUATT-3'; USP7, 5'-GGCGAAGUUUAAUGUAUUTT-3'; and USP9x, 5'-GCAGUGAGUGG CUGGAAGUTT-3'.

**Cell culture.** Human U2OS, HCT116 and RPE1 cells were obtained from ATCC. U2OS and HCT116 were cultured in DMEM containing 10% FBS and 1×penicillin-streptomycin, while RPE1 cells were grown in a 1:1 mixture of Ham's F12 and DMEM supplemented with 10% FBS and 1×penicillin-streptomycin. Serum-starvation of RPE1 cells was done by incubating cells for 24 h in medium supplemented with 0.25% FBS. A HCT116 UBC13-knockout cell line was generated using CRISPR-Cas9 technology<sup>14,15</sup>. A donor plasmid bearing a splice acceptor site and a puromycin resistance marker, flanked by homology arms, was co-transfected with pX300 (ref. 14) targeting the GGCGCGCGGAATCGCGGCG sequence within the first intron of the UBC13 gene. To generate cell lines capable of doxycycline-induced expression of GFP-tagged K63-Super-UIM, U2OS cells were transfected with GFP-K63-Super-UIM plasmid and pcDNA6/TR and positive clones were selected with Zeocin (Invitrogen) and Blasticidin S (Invitrogen). Stable U2OS cell lines expressing RNF8 or RNF168 shRNA in a doxycycline-inducible manner or Strep-HA-ubiquitin were described previously<sup>1,4,31</sup>. All cell lines were regularly tested for mycoplasma infection. Unless otherwise indicated, cells were exposed to DSBs using IR (4 Gy for microscopy experiments and 10 Gy for biochemical analyses) or laser micro-irradiation (as described previously<sup>32</sup>), and collected 1 h later.

**Recombinant protein production.** Purified biotinylated K63-Super-UIM wild-type and mutant proteins containing an N-terminal, biotinylated BioEase tag and a C-terminal His<sub>6</sub>-tag were obtained by expressing the proteins in an *E. coli* strain expressing the BirA biotin ligase. Bacteria were grown in LB medium containing 0.5 mM biotin, induced with 0.25 mM isopropyl-β-D-thiogalactoside (IPTG) for 3 h at 30 °C, and then lysed by French press. The K63-Super-UIM constructs were purified using immobilized metal affinity chromatography (IMAC) followed by size-exclusion chromatography (SEC). Purity and complete biotinylation of the proteins was verified by mass spectrometry. Recombinant Strep-His<sub>6</sub>-RNF168

UDM-1/2 was produced in Rosetta2(DE3)pLacI (Novagen) bacteria induced with 0.5 mM IPTG for 3 h at 30 °C, lysed using Bugbuster (Novagen) supplemented with Protease Inhibitor Cocktail without EDTA (Roche). The proteins were purified on Ni<sup>2+</sup>-NTA-agarose (Qiagen). Recombinant human UBA1, UBCH5c, UBC13, MMS2, RNF8 and ubiquitin used for *in vitro* ubiquitylation assays were purified as described<sup>8</sup>.

**Antibodies.** Antibodies used in this study included: UBC13 (#4919, Cell Signaling), MCM6 (sc-9843, Santa Cruz), 53BP1 (sc-22760, Santa Cruz), γ-H2A.X (05-636, Millipore; or 2577, Cell Signaling), H2A.X (2595, Cell Signaling), MDC1 (ab11171, Abcam), conjugated ubiquitin (FK2) (BML-PW8810-0500, Enzo Life Sciences), HA (11867423991, Roche; and sc-7392, Santa Cruz), Myc (sc-40, Santa Cruz), His<sub>6</sub> (631212, Clontech), GFP (sc-9996, Santa Cruz; 11814460001, Roche), ubiquitin (sc-8017, Santa Cruz), histone H1.2 (ab17677, Abcam), histone H1x (A304-604A, Bethyl Labs), histone H1 (pan, #AE-4 clone) (ab7789, Abcam), histone H2A (07-146, Millipore), histone H2B (ab1790, Abcam), histone H3 (ab1791, Abcam), histone H4 (ab7311, Abcam), cyclin A (sc-751, Santa Cruz), actin (MAB1501, Millipore), BRCA1 (sc-6954, Santa Cruz), RNF168 for immunofluorescence (06-1130, Millipore) and antibody to RNF168 (ref. 5) used for immunoblots were gifts from D. Durocher. Antibody to RNF8 has been described previously<sup>1</sup>.

**Immunochemical methods.** For pull-down of K63-ubiquitylated proteins, cells were lysed in high-stringency buffer (50 mM Tris, pH 7.5; 500 mM NaCl; 5 mM EDTA; 1% NP40; 1 mM dithiothreitol (DTT); 0.1% SDS) containing 1.25 mg ml<sup>-1</sup> N-ethylmaleimide, 50 μM DUB inhibitor PR619 (LifeSensors), and protease inhibitor cocktail (Roche). Recombinant biotinylated K63-Super-UIM (25 μg ml<sup>-1</sup>) was added immediately upon lysis, followed by sonication and centrifugation. Streptavidin M-280 Dynabeads (Invitrogen) was added to immobilize the K63-Super-UIM, and bound material was washed extensively in high-stringency buffer. A Benzonase (Sigma) and MNase (NEB) treatment step was included to remove any contaminating nucleotides. Proteins were resolved by SDS-PAGE and analysed by immunoblotting. Where indicated, bound complexes were subjected to deubiquitylation by incubation with USP2c (1 μM, Boston Biochem) in DUB buffer (50 mM HEPES, pH 7.5; 100 mM NaCl; 1 mM MnCl<sub>2</sub>; 0.01% Brij-35; 2 mM DTT) overnight at 30 °C before boiling in Laemmli Sample Buffer. Immunoblotting, Strep-Tactin pull-downs, and chromatin enrichment were done essentially as described<sup>32</sup>. Briefly, Strep-RNF168 UDM pull-down experiments from cells were performed after lysing cells in EBC buffer (50 mM Tris, pH 7.4; 150 mM NaCl; 0.5% NP-40; 1 mM EDTA) containing 1.25 mg ml<sup>-1</sup> NEM, 50 μM PR619 (LifeSensors) and protease inhibitor cocktail (Roche). The soluble fraction was subsequently used for immunoprecipitation using Strep-Tactin sepharose (IBA). After washing in EBC buffer, proteins were eluted and analysed by immunoblotting. To isolate Strep-HA-ubiquitin-conjugated proteins, cells were lysed in denaturing buffer (20 mM Tris, pH 7.5; 50 mM NaCl; 1 mM EDTA; 1 mM DTT; 0.5% NP-40; 0.5% sodium deoxycholate; 0.5% SDS) containing 1.25 mg ml<sup>-1</sup> NEM, 50 μM PR619 (LifeSensors) and protease inhibitor cocktail (Roche). After sonication and centrifugation, Strep-HA-ubiquitin-conjugated proteins were immobilized on Strep-Tactin sepharose (IBA). After extensive washing in denaturing buffer, proteins were eluted and analysed by immunoblotting. For chromatin fractionation, cells were first lysed in buffer 1 (100 mM NaCl; 300 mM sucrose; 3 mM MgCl<sub>2</sub>; 10 mM PIPES, pH 6.8; 1 mM EGTA; 0.2% Triton X-100) containing protease, phosphatase and DUB inhibitors and incubated on ice for 5 min. After centrifugation, the soluble proteins were removed and the pellet was resuspended in buffer 2 (50 mM Tris-HCl, pH 7.5; 150 mM NaCl; 5 mM EDTA; 1% Triton X-100; 0.1% SDS) containing protease, phosphatase and DUB inhibitors. Lysates were then incubated 10 min on ice, sonicated, and solubilized chromatin-enriched fractions were collected after centrifugation.

**Immunofluorescence staining and microscopy.** For immunofluorescence staining, cells were fixed in 4% paraformaldehyde for 15 min, permeabilized with PBS containing 0.2% Triton X-100 for 5 min, and incubated with primary antibodies diluted in DMEM for 1 h at room temperature. After staining with secondary antibodies (Alexa Fluor; Life Technologies) for 1 h, coverslips were mounted in Vectashield mounting medium (Vector Laboratories) containing nuclear stain DAPI. Images of GFP-K63-Super-UIM were all obtained from a stable cell line where GFP-K63-Super-UIM was induced by incubating with 1 μg ml<sup>-1</sup> doxycycline for approximately 24 h unless otherwise stated. Images were acquired with an LSM 780 confocal microscope (Carl Zeiss Microimaging) mounted on Zeiss-Axiovert 100M equipped with Plan-Apochromat 40×/1.3 oil immersion objective, using standard settings. Image acquisition and analysis was carried out with ZEN2010 software. For ImageJ-based image analysis, images were acquired with an AF6000 wide-field microscope (Leica Microsystems) equipped with a Plan-Apochromat 40×/0.85 CORR objective, using the same microscopic settings. Fluorescence intensities of the micro-irradiated region (demarcated by γ-H2AX positivity) and the nucleus were first corrected for the general image background.

Using these values, relative recruitment to DNA damage sites (relative fluorescence units (RFUs)) was calculated by normalizing the nuclear-background-corrected signal at the micro-irradiated region to that of the nuclear background. Finally, the RFU of the protein of interest was normalized to the RFU of the  $\gamma$ -H2AX signal and plotted as the average of biological triplicates. Fluorescence recovery after photobleaching (FRAP) was performed essentially as described<sup>33</sup>. Briefly, U2OS cells stably expressing GFP-H1 were grown in glass-bottom dishes (LabTek) in the presence of CO<sub>2</sub>-independent medium. A 2- $\mu$ m-wide rectangular strip spanning the entire width of the cell was bleached by excitation with the maximal intensity of a 488 nm laser line, after which 95 frames of the bleached region were acquired at 4 s intervals. Mean fluorescence intensities were processed, normalized and analysed as described<sup>33</sup>.

**In vitro binding and ubiquitylation assays.** Binding of K63-Super-UIM to di-ubiquitin (Ub<sub>2</sub>) linkages (Boston Biochem) was done by incubating 100 ng Ub<sub>2</sub> with 2.5  $\mu$ g K63-Super-UIM immobilized on Streptavidin M-280 Dynabeads (Invitrogen) in buffer A (50 mM Tris, pH 7.5; 10% glycerol; 400 mM NaCl; 0.5% NP40; 2 mM DTT; 0.1 mg ml<sup>-1</sup> BSA). After extensive washing, bound complexes were resolved by SDS-PAGE and analysed by immunoblotting. Binding of RNF168 UDM1/2 to di-ubiquitin (Ub<sub>2</sub>) linkages was analysed by incubating 100 ng Ub<sub>2</sub> with 5  $\mu$ g Strep-RNF168-UDM1/2 immobilized on Strep-Tactin sepharose (IBA BioTAGnology) in buffer B (50 mM Tris, pH 8; 5% glycerol; 0.5% NP40; 2 mM DTT; 0.1 mg ml<sup>-1</sup> BSA; 2 mM MgCl<sub>2</sub>, supplemented with 250 mM KCl for UDM1 binding and 100 mM KCl for UDM2 binding). After extensive washing, bound complexes were resolved by SDS-PAGE and analysed by immunoblotting. Where indicated, UDM1/2 binding to K63-linked Ub<sub>2</sub> was analysed in the presence of increasing KCl concentrations (75 mM, 150 mM and 250 mM). To analyse binding of RNF168 UDM1/2 to recombinant histones, purified Strep-RNF168 UDM1/2 (10  $\mu$ g) was pre-bound to Strep-Tactin sepharose in buffer C (for binding to H1.0) (50 mM, Tris pH 8; 5% glycerol; 150 mM KCl; 0.5% NP40; 2 mM DTT; 0.1 mg ml<sup>-1</sup> BSA) or D (for binding to H2A) (50 mM, Tris pH 8; 5% glycerol; 75 mM KCl; 0.05% NP40; 2 mM DTT; 0.1 mg ml<sup>-1</sup> BSA), and incubated with 500 ng recombinant histone H1.0 or H2A (New England Biolabs). Bound complexes were washed and analysed by immunoblotting. To analyse binding of LRM1 and LRM2 peptides to histone H1.0 or H2A, magnetic Streptavidin beads were incubated with buffer E (25 mM, Tris pH 8.5; 5% glycerol; 50 mM KCl; 0.5% TX-100; 1 mM DTT; 0.1 mg ml<sup>-1</sup> BSA) in the absence (control) or presence of 1.5  $\mu$ g purified, biotinylated RNF168 LRM1 (amino acids 110–133) or LRM2 (amino acids 463–485) peptide. Samples were then incubated with 250 ng recombinant H2A or H1.0 for 2 h at 4 °C, and immobilized complexes were washed and analysed by SDS-PAGE and Colloidal Blue staining (Invitrogen).

For *in vitro* ubiquitylation assays, histone-H1-containing oligonucleosomes (10  $\mu$ M) were purified in the presence of 55 mM iodoacetamide, essentially as described previously<sup>34</sup>, with the exception that micrococcal nuclease digestion was stopped with 20 mM EGTA and dialysis was started right after the second homogenization in buffer containing 50 mM Tris, pH 7.5; 150 mM NaCl; 1 mM TCEP; and 340 mM sucrose. Dialysed samples were then incubated with DUB inhibitor (Ubiquitin-PA<sup>35</sup>, 20  $\mu$ M) for 20 min at room temperature. Nucleosomes were incubated with 0.5  $\mu$ M human UBA1, 5  $\mu$ M UBCH5c, 1  $\mu$ M UBC13-MMS2 complex, 5  $\mu$ M RNF8<sub>351–485</sub> fragment (purified as described previously<sup>36</sup>) and 75  $\mu$ M ubiquitin in reaction buffer (50 mM Tris, pH 7.5; 100 mM NaCl; 3 mM ATP; 3 mM MgCl<sub>2</sub>; 1 mM TCEP) at 31 °C. Samples were analysed by immunoblot analysis.

**SILAC-based quantification of di-glycine-containing peptides.** For SILAC experiments, U2OS or HCT116 cells were grown in medium containing unlabelled L-arginine and L-lysine (Arg<sup>0</sup>/Lys<sup>0</sup>) as the light condition, or isotope-labelled variants of L-arginine and L-lysine (Arg<sup>6</sup>/Lys<sup>4</sup> or Arg<sup>10</sup>/Lys<sup>8</sup>) as the heavy condition<sup>36</sup>. SILAC-labelled HCT116 wild-type and UBC13-knockout cells were lysed in modified RIPA buffer (50 mM Tris-HCl, pH 7.5; 150 mM NaCl; 1% Nonidet P-40; 0.1% sodium-deoxycholate; 1 mM EDTA) supplemented with protease inhibitors (complete protease inhibitor mixture tablets, Roche Diagnostics) and N-ethylmaleimide (5 mM). Lysates were incubated for 10 min on ice and cleared by centrifugation at 16,000g. An equal amount of protein from the two SILAC states was mixed and precipitated by adding fivefold acetone and incubating at -20 °C overnight. Precipitated proteins were dissolved in denaturing buffer (6 M urea; 2 M thiourea; 10 mM HEPES, pH 8.0), reduced with DTT (1 mM) and alkylated with chloroacetamide (5.5 mM). Proteins were digested with lysyl endoproteinase C (Lys-C) for 6 h, diluted fourfold with water and digested overnight with trypsin. The digestion was stopped by addition of trifluoroacetic acid (0.5% final concentration), incubated at 4 °C for 2 h and centrifuged for 15 min at 4,000g. Peptides from the cleared solution were purified by reversed-phase Sep-Pak C18 cartridges (Waters Corporation). Diglycine-lysine modified peptides were enriched using the Ubiquitin Remnant Motif Kit (Cell Signaling

Technology), according to the manufacturer's instructions. Briefly, peptides were eluted from the Sep-Pak C18 cartridges with 50% acetonitrile, which was subsequently removed by centrifugal evaporation. Peptides were incubated with 40  $\mu$ l of anti-di-glycine-lysine antibody resin in immunoaffinity purification (IAP) buffer for 4 h at 4 °C. Beads were washed three times with IAP buffer, two times with water and peptides eluted with 0.15% trifluoroacetic acid. Eluted peptides were fractionated by microcolumn-based strong cation exchange chromatography (SCX) and cleaned by reversed-phase C18 stage-tips.

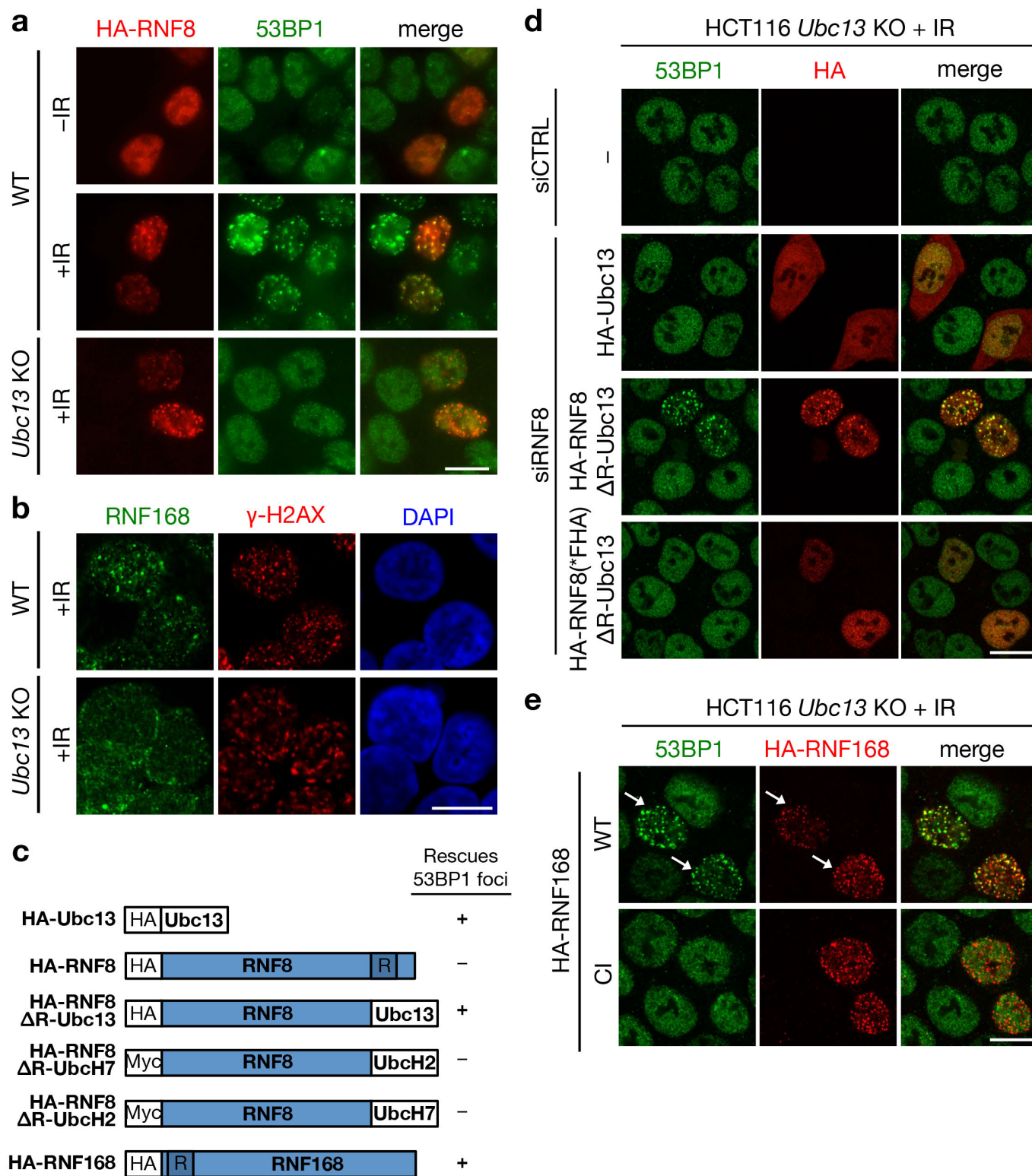
**SILAC K63-Super-UIM pull-down and in-gel digestion.** SILAC-labelled cells were lysed in high-stringency RIPA buffer (50 mM Tris-HCl, pH 7.5; 500 mM NaCl; 1% Nonidet P-40; 0.1% sodium-deoxycholate; 1 mM EDTA) containing 1.25 mg ml<sup>-1</sup> N-ethylmaleimide, 50  $\mu$ M DUB inhibitor PR619 (LifeSensors), and protease inhibitor cocktail (Roche). Lysates from different SILAC states were separately incubated for 10 min on ice and cleared by centrifugation at 16,000g. Extracts (5 mg) were incubated for 4 h at 4 °C with K63-Super-UIM immobilized to Streptavidin beads (approximately 5  $\mu$ g K63-Super-UIM per experiment). Beads were washed three times with high-stringency RIPA, beads from the different SILAC conditions were mixed, and proteins were eluted with SDS sample buffer, incubated with DTT (10 mM) for 10 min at 70 °C and alkylated with chloroacetamide (5.5 mM) for 60 min at 25 °C. Proteins were separated by SDS-PAGE using a 4–12% gradient gel and visualized with colloidal blue stain. Gel lanes were sliced into six pieces, and proteins were digested in-gel using standard methods<sup>37</sup>.

**Mass spectrometry and data analysis.** Peptides were analysed on a quadrupole Orbitrap (Q Exactive, Thermo Scientific) mass spectrometer equipped with a nanoflow HPLC system (Thermo Scientific). Peptide samples were loaded onto C18 reversed-phase columns and eluted with a linear gradient (1–2 h for in-gel samples, and 3–4 h for di-glycine-lysine enriched samples) from 8 to 40% acetonitrile containing 0.5% acetic acid. The mass spectrometer was operated in a data-dependent mode automatically switching between MS and MS/MS. Survey full scan MS spectra ( $m/z$  300–1200) were acquired in the Orbitrap mass analyser. The 10 most intense ions were sequentially isolated and fragmented by higher-energy C-trap dissociation (HCD). Peptides with unassigned charge states, as well as peptides with charge state less than +2 for in-gel samples and +3 for di-glycine-lysine enriched samples were excluded from fragmentation. Fragment spectra were acquired in the Orbitrap mass analyser. Raw MS data were analysed using MaxQuant software (version 1.3.9.21). Parent ion and tandem mass spectra were searched against protein sequences from the UniProt knowledge database using the Andromeda search engine. Spectra were searched with a mass tolerance of 6 ppm in the MS mode, 20 ppm for MS/MS mode, strict trypsin specificity and allowing up to two missed cleavage sites. Cysteine carbamidomethylation was searched as a fixed modification, whereas amino-terminal protein acetylation, methionine oxidation and N-ethylmaleimide modification of cysteines, and di-glycine-lysine were searched as variable modifications. Di-glycine-lysines were required to be located internally in the peptide sequence. Site localization probabilities were determined using MaxQuant (PTM scoring algorithm) as described previously<sup>38</sup>. A false discovery rate of less than 1% was achieved using the target-decoy search strategy<sup>39</sup> and a posterior error probability filter. Information about previously known protein–protein interactions among putative UBC13-dependent K63-Super-UIM interacting proteins was extracted using the HIPPIE database<sup>40</sup> (version 1.6), and interactions were visualized in Cytoscape<sup>41</sup>. The Gene Ontology (GO) biological process term analysis for UBC13-dependent K63-Super-UIM interacting proteins was filtered for categories annotated with at least 20 and not more than 300 genes. Redundant GO terms (less than 30% unique positive-scoring genes compared to more significant GO term) were removed and the five most significant (Fisher's exact *t*-test) remaining GO term categories depicted. To determine the variation within the quantification of ubiquitin linkage types, an F-test was performed and the *P* values were adjusted using the Bonferroni method. A significant difference in the variances between K48 and K11, and K48 and K6 ubiquitin linkages was detected. To test the significance of the difference between the SILAC ratios measured for ubiquitin linkage types, the Welch two-sample *t*-test was performed and the obtained *P* values were adjusted using the Bonferroni method.

- Danielsen, J. M. *et al.* Mass spectrometric analysis of lysine ubiquitylation reveals promiscuity at site level. *Mol. Cell. Proteomics* **10**, M110.003590 (2011).
- Poulsen, M., Lukas, C., Lukas, J., Bekker-Jensen, S. & Mailand, N. Human RNF169 is a negative regulator of the ubiquitin-dependent response to DNA double-strand breaks. *J. Cell Biol.* **197**, 189–199 (2012).
- Bekker-Jensen, S., Lukas, C., Melander, F., Bartek, J. & Lukas, J. Dynamic assembly and sustained retention of 53BP1 at the sites of DNA damage are controlled by Mdc1/NFBD1. *J. Cell Biol.* **170**, 201–211 (2005).

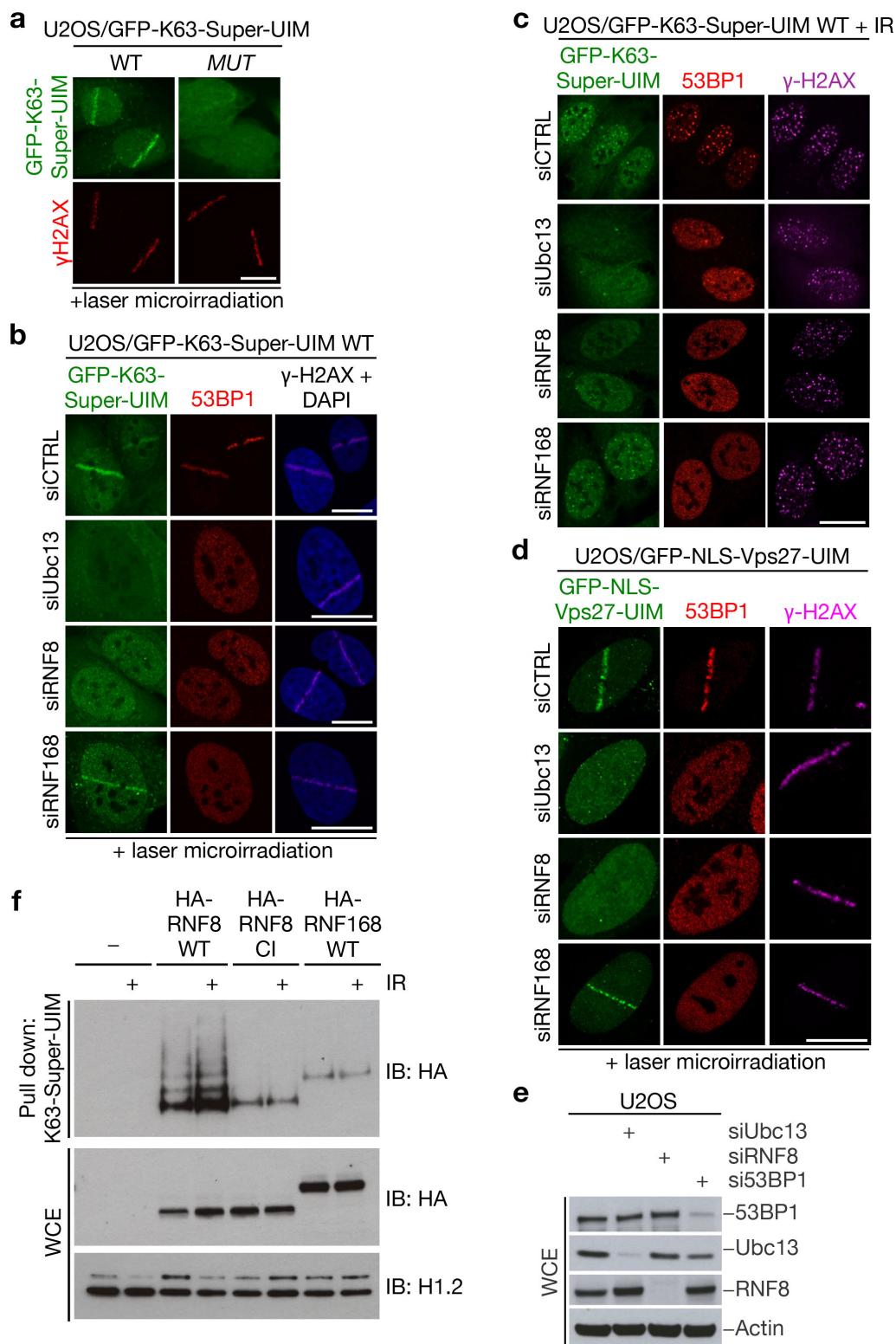
34. Hernández-Muñoz, I. *et al.* Stable X chromosome inactivation involves the PRC1 Polycomb complex and requires histone MACROH2A1 and the CULLIN3/SPOP ubiquitin E3 ligase. *Proc. Natl Acad. Sci. USA* **102**, 7635–7640 (2005).
35. Ekkebus, R. *et al.* On terminal alkynes that can react with active-site cysteine nucleophiles in proteases. *J. Am. Chem. Soc.* **135**, 2867–2870 (2013).
36. Ong, S. E. *et al.* Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1**, 376–386 (2002).
37. Jensen, O. N., Wilm, M., Shevchenko, A. & Mann, M. Sample preparation methods for mass spectrometric peptide mapping directly from 2-DE gels. *Methods Mol. Biol.* **112**, 513–530 (1999).
38. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnol.* **26**, 1367–1372 (2008).
39. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods* **4**, 207–214 (2007).
40. Schaefer, M. H. *et al.* HIPPIE: Integrating protein interaction networks with experiment based quality scores. *PLoS ONE* **7**, e31826 (2012).
41. Cline, M. S. *et al.* Integration of biological networks and gene expression data using Cytoscape. *Nature Protocols* **2**, 2366–2382 (2007).





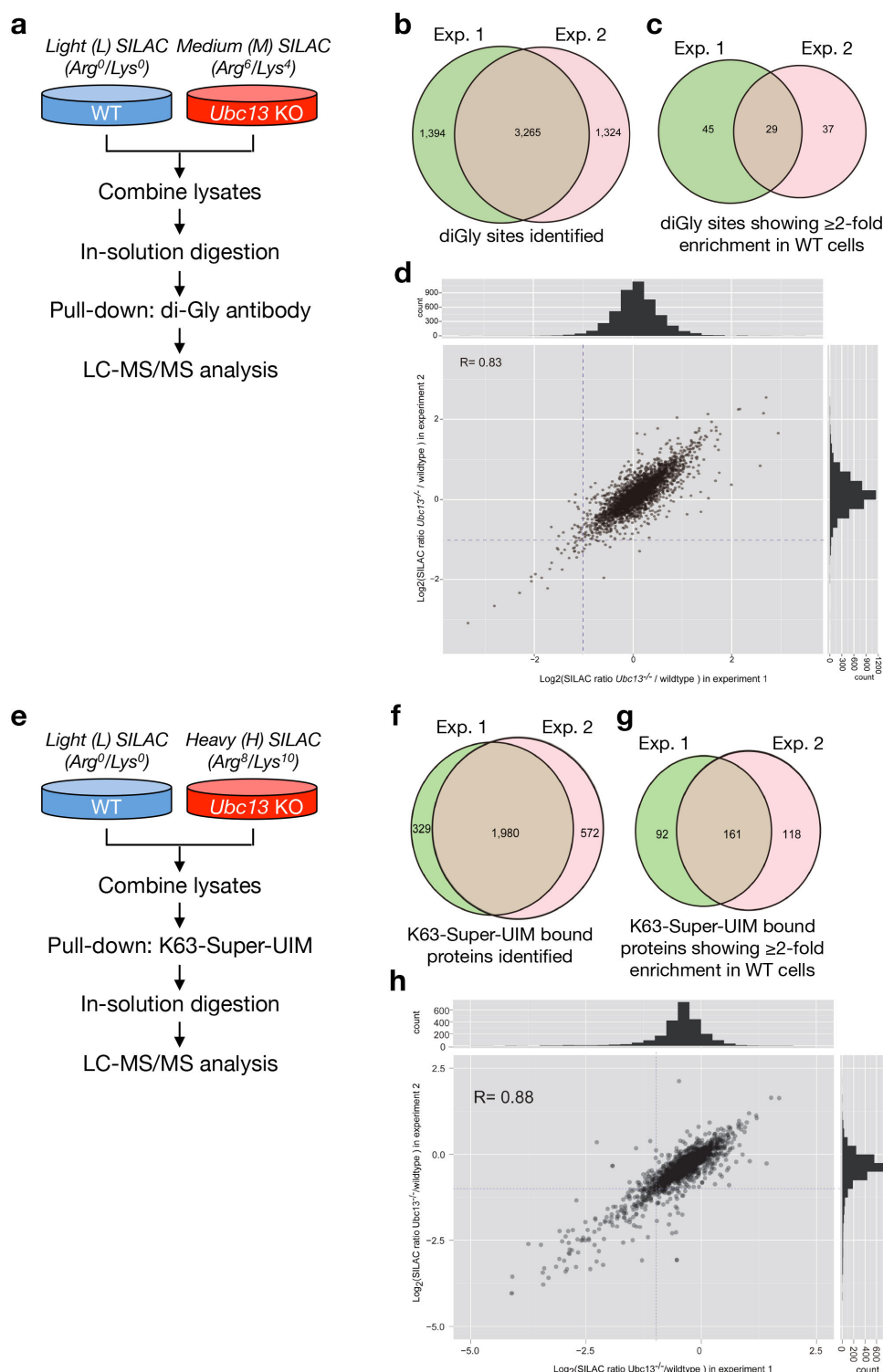
**Extended Data Figure 1 | Protein recruitment to DSB sites in *UBC13*-knockout cells.** **a, b**, Representative images of HCT116 wild-type (WT) or *UBC13*-knockout (KO) cells exposed to IR. Where indicated, cells were transfected with HA-RNF8 plasmid before IR ( $n = 2$  experiments). **c**, Constructs used in Fig. 1d and their ability to restore IR-induced 53BP1 foci in *UBC13*-knockout cells. **d**, Representative images of HCT116 *UBC13*-knockout cells transfected with non-targeting control (CTRL) or RNF8

siRNAs and subsequently with plasmids encoding *UBC13* or siRNA-resistant *UBC13*-RNF8 fusion constructs ( $n = 2$ ). **e**, Representative images of *UBC13*-knockout cells transfected with plasmids encoding wild-type or catalytically inactive (CI) HA-RNF168 ( $n = 2$ ). Expression of HA-RNF168 wild type restores IR-induced 53BP1 foci formation (arrows). Scale bars, 10  $\mu$ m.



**Extended Data Figure 2 | RNF8- and UBC13-dependent K63-linked ubiquitylation at DSB sites.** **a–c**, Representative images of U2OS cells stably expressing GFP-K63-Super-UIM, transfected with siRNAs where indicated, and exposed to laser micro-irradiation or IR ( $n = 3$ ). **d**, Representative images of U2OS cells stably expressing GFP- and nuclear localization signal (NLS)-tagged Vps27-UIM (with high affinity for binding to K63-linked ubiquitin<sup>18</sup>) transfected with the indicated siRNAs and exposed to laser micro-irradiation ( $n = 2$ ). **e**, Loss of RNF8 or UBC13 has no impact on 53BP1 abundance.

Immunoblot analysis of whole-cell extracts (WCEs) of U2OS cells transfected with indicated siRNAs. **f**, RNF8, but not RNF168, is modified by K63-linked ubiquitin chains. K63-Super-UIM pull-downs from U2OS cells transfected with empty vector (–), HA-tagged RNF8 (wild type (WT) or catalytically inactive (CI)) or HA-RNF168 plasmids were immunoblotted (IB) with indicated antibodies. Scale bars, 10  $\mu$ m. **e**, **f**, Uncropped blots are shown in Supplementary Fig. 1.

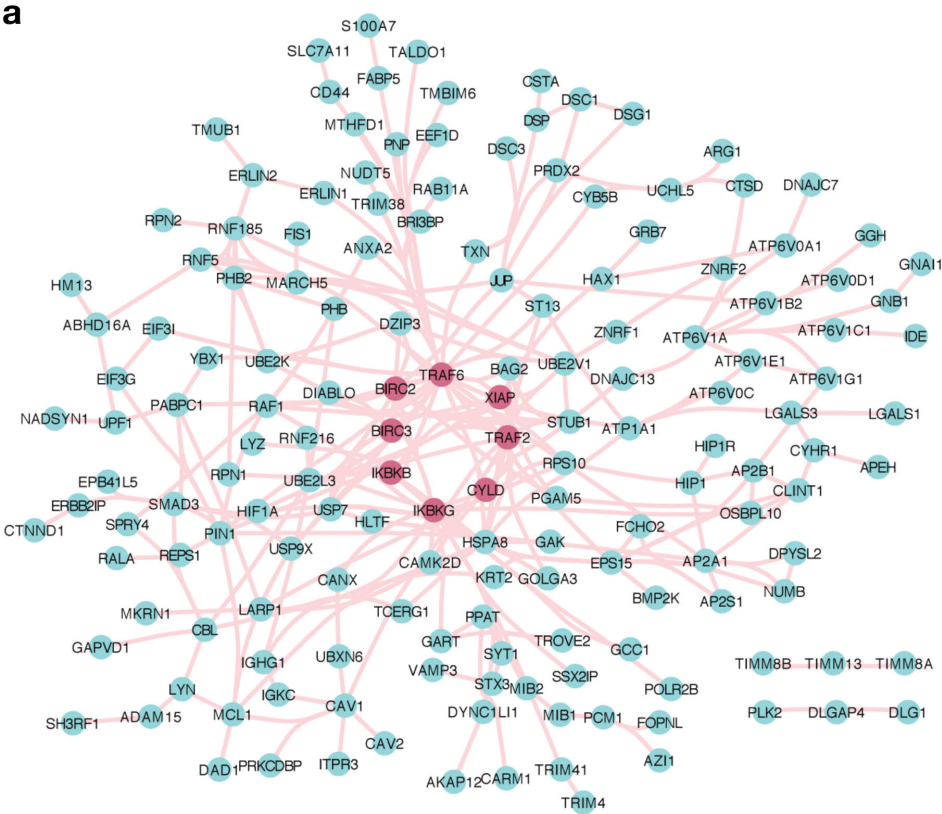


**Extended Data Figure 3 | Experimental replicates of SILAC-based quantification of di-glycine-containing peptides and K63-Super-UIM pull-downs from wild-type and *Ubc13*-knockout cells.** **a**, Schematic outline of SILAC-based mass spectrometry approach to quantify di-glycine-containing peptides in HCT116 wild-type (WT) and *Ubc13*-knockout (KO) cells. **b**, **c**, Proportional Venn diagrams showing overlap between all identified di-glycine-containing peptides (**b**) and those with a SILAC ratio (*Ubc13*-knockout/wild-type cells)  $< 0.5$  (**c**) in two independent experiments (Exp.) performed as shown in **a** (Supplementary Table 1). **d**, Scatter plot showing correlation between SILAC ratios of di-glycine-containing peptides. The

Pearson's correlation coefficient (*R*) is indicated. **e**, Schematic outline of SILAC-based mass spectrometry approach to identify *Ubc13*-dependent K63-ubiquitylation targets in unperturbed HCT116 wild-type and *Ubc13*-knockout cells. **f**, **g**, Proportional Venn diagrams showing overlap between all proteins identified in K63-Super-UIM pull-downs (**f**) and those with a SILAC ratio (*Ubc13*-knockout/wild-type cells)  $< 0.5$  (**g**) in two independent experiments performed as shown in **e** (Supplementary Table 2). **h**, Scatter plot showing correlation between SILAC ratios of proteins identified in two experiments. The Pearson's correlation coefficient (*R*) is indicated.

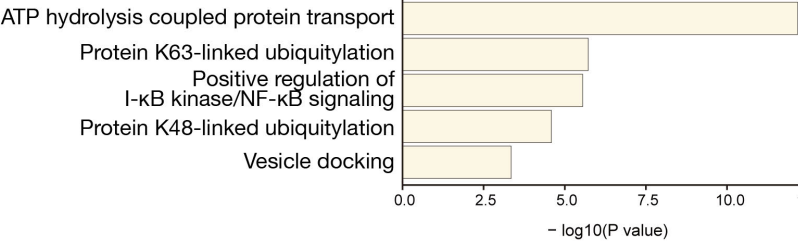


**a**

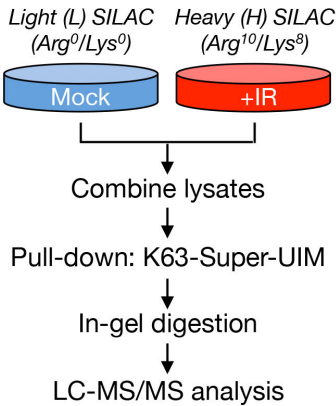


**b**

GO biological process



**c**

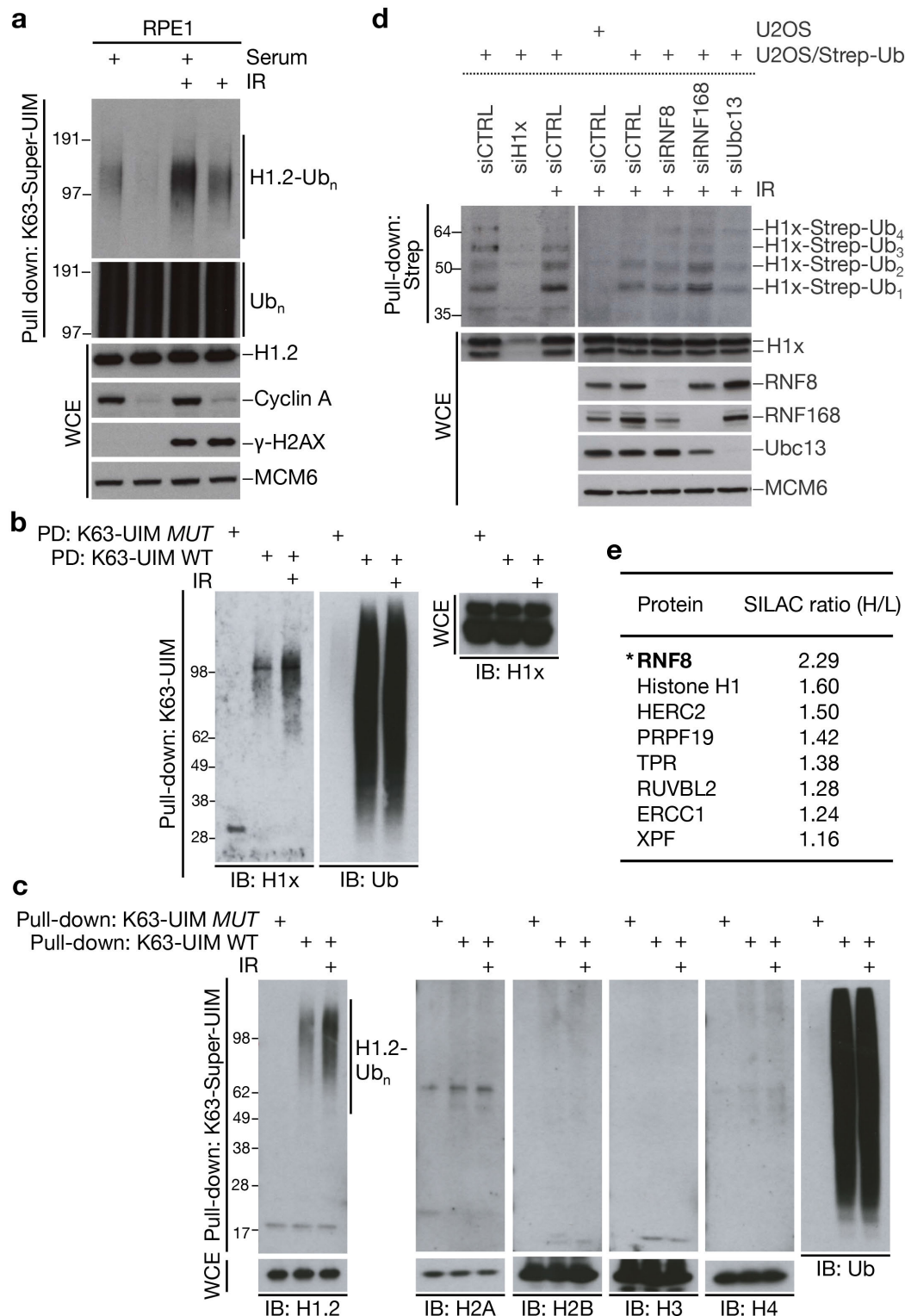


**d**

Protein	SILAC ratio (H/L)
TRAF6	3.41
TRAF2	2.91
HP1BP3	2.62
<b>Histone H1.2</b>	<b>1.85</b>
<b>Histone H1x</b>	<b>1.66</b>
<b>Histone H1.0</b>	<b>1.40</b>
CENPF	1.36
SUMO1	1.36
Histone H2A	1.10
Histone H3	1.09
Histone H2B	0.96
Histone H4	0.88

**Extended Data Figure 4 | Analysis of UBC13-dependent K63-ubiquitylated proteins in unperturbed cells and in response to DNA damage.** **a**, K63 linkages from extracts of HCT116 wild-type (WT) and *UBC13*-knockout (KO) cells were enriched by K63-Super-UIM pull-down (Supplementary Table 2). Interaction network shows proteins enriched at least twofold in wild-type cells. Proteins involved in UBC13-dependent activation of NF- $\kappa$ B signalling are

highlighted in red. **b**, Functional annotation of potential UBC13-dependent K63-ubiquitylated proteins (**a**), showing enriched Gene Ontology (GO) biological process terms. **c**, Schematic outline of SILAC-based mass spectrometry approach to identify targets of K63 ubiquitylation in response to IR-induced DSBs. **d**, SILAC ratios of selected proteins from U2OS cells treated as in **c**. Data from a representative experiment are shown ( $n = 3$ ).

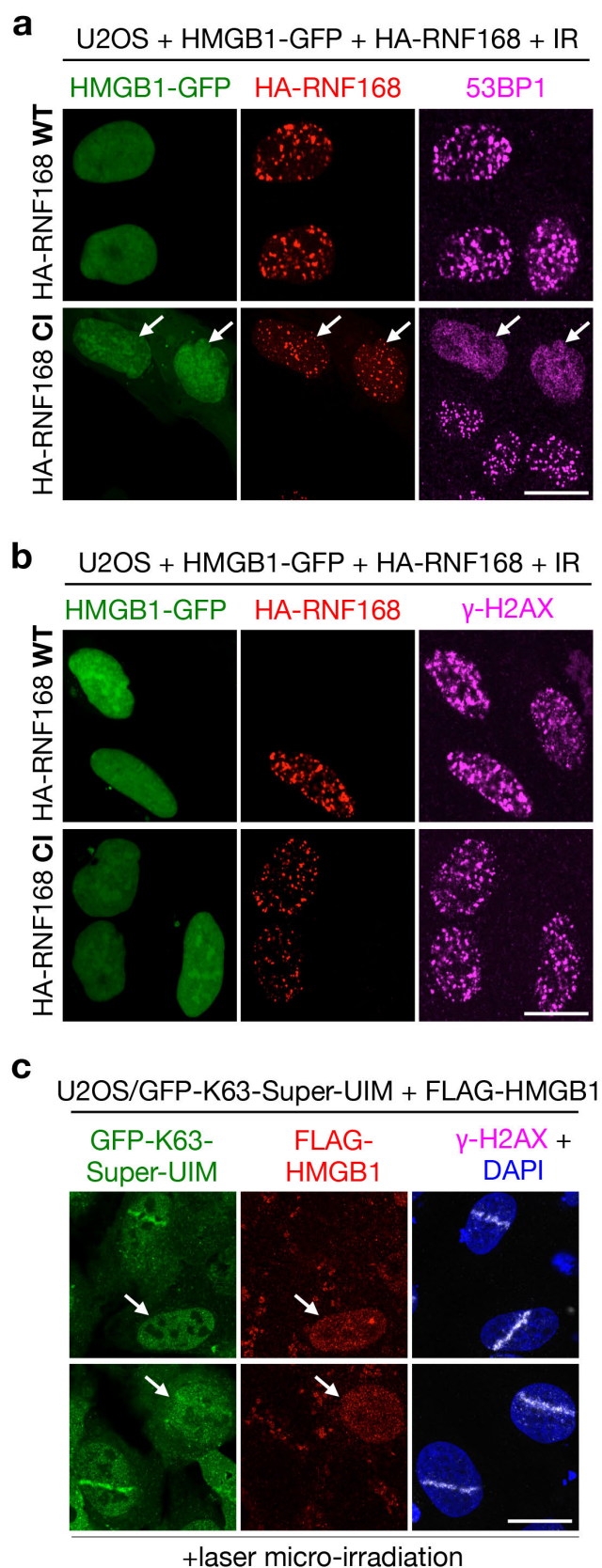


**Extended Data Figure 5 | DSB-induced K63 ubiquitylation of H1-type linker histones.** **a**, Analysis of K63-linked ubiquitylation of histone H1.2 in RPE1 cells growing exponentially or kept quiescent by serum starvation. **b**, **c**, K63-Super-UIM pull-downs from U2OS cells exposed or not to IR were immunoblotted (IB) with the indicated antibodies. **d**, U2OS cells or U2OS cells stably expressing Strep-HA-ubiquitin (U2OS/Strep-Ub) were transfected with the indicated siRNAs and exposed or not to IR. Whole-cell extracts (WCEs) and Strep-ubiquitin-conjugated proteins immobilized on Strep-Tactin beads under denaturing conditions were analysed by immunoblotting.

**e**, Proteins interacting with endogenous RNF8. U2OS cells stably expressing RNF8 shRNA in a doxycycline (DOX)-inducible manner<sup>1</sup> was grown in light (L) or heavy (H) SILAC medium. Cells growing in light medium were induced to express RNF8 shRNA by treatment with DOX. Both cultures were then exposed to IR and processed for immunoprecipitation with RNF8 antibody. Bound proteins were analysed by mass spectrometry. Proteins displaying the highest H/L SILAC ratios are listed. **a–d**, The migration of molecular weight markers (kDa) is indicated on the left. Uncropped blots are shown in Supplementary Fig. 1.

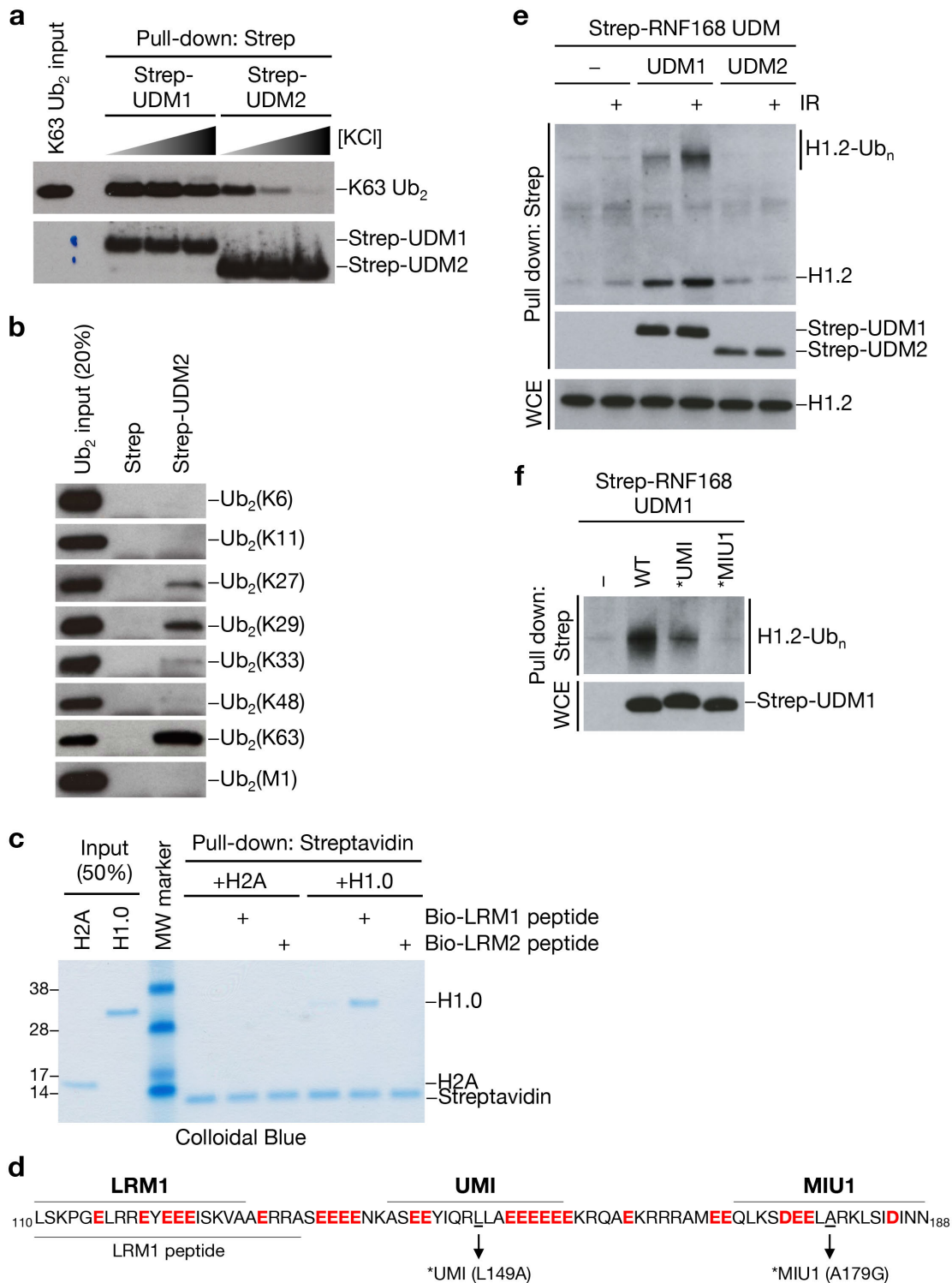






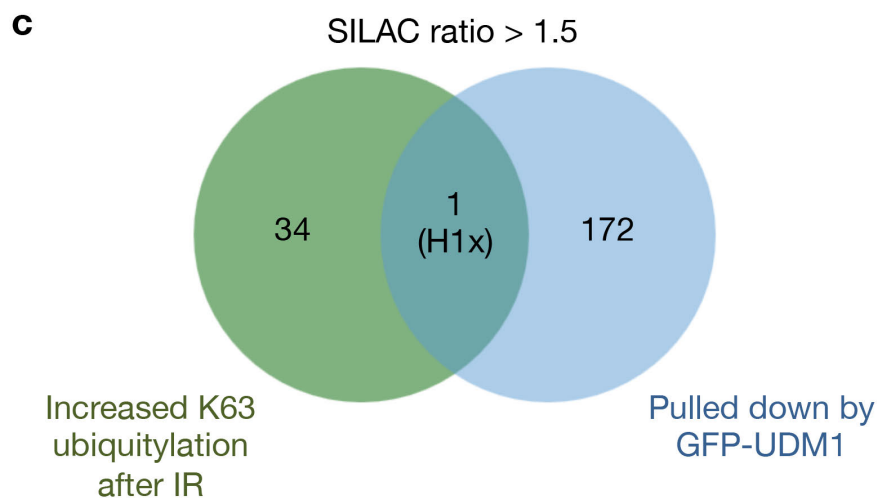
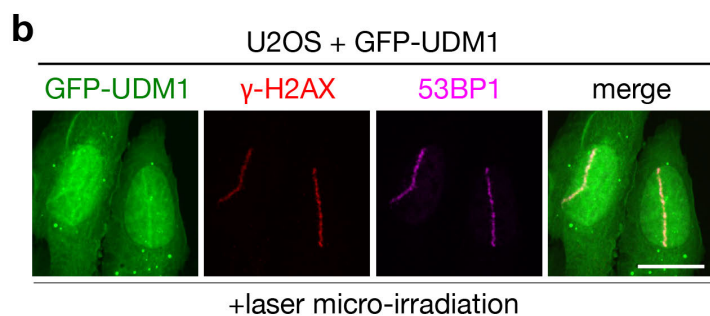
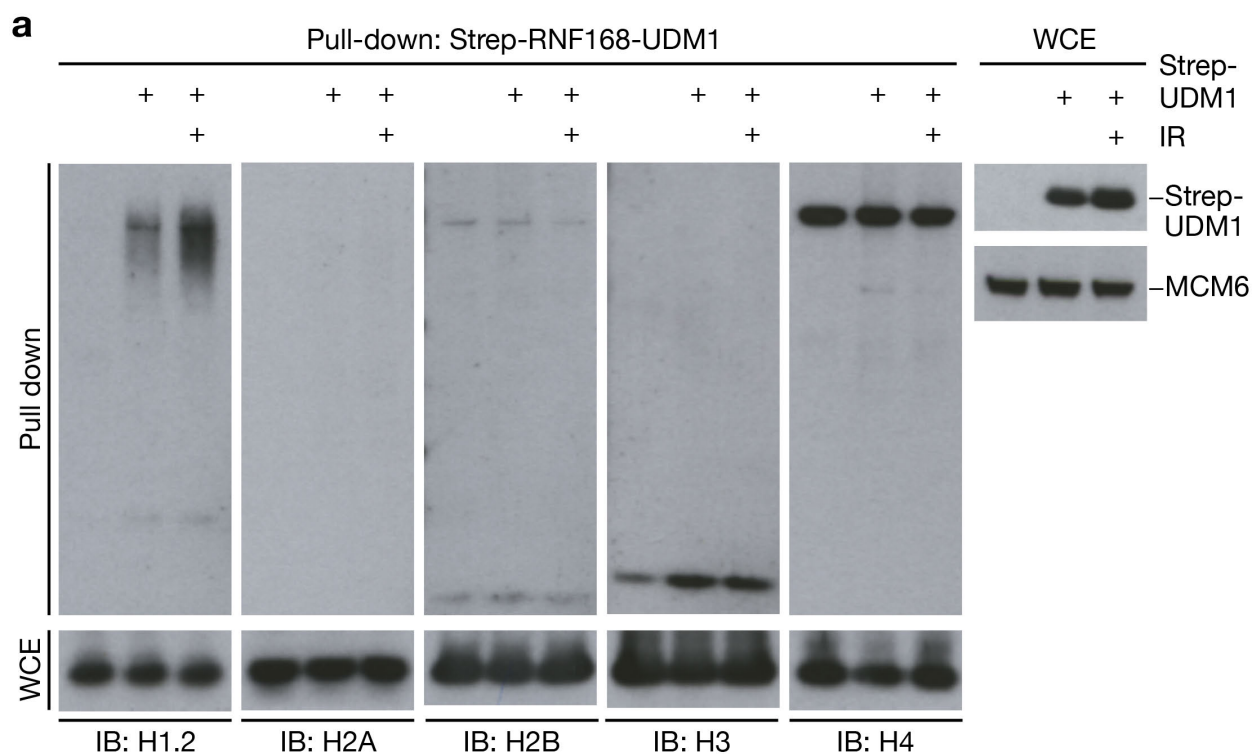
**Extended Data Figure 7 | HMGB1 overexpression impairs the RNF8/RNF168-dependent signalling response at DSB sites at the level of K63 ubiquitylation and RNF168 recruitment.** a, b, Representative images of U2OS cells co-transfected with constructs encoding HMGB1-GFP and wild-type (WT) or catalytically inactive (CI) HA-RNF168 and exposed to IR

( $n = 3$ ). Arrows indicate cells expressing HA-RNF168 CI, in which 53BP1 foci formation is not restored. c, Representative images of U2OS/GFP-K63-Super-UIM cells transfected with Flag-HMGB1 construct and subjected to laser micro-irradiation ( $n = 3$ ). Flag-HMGB1-expressing cells show reduced K63 ubiquitylation at DSB sites (indicated by arrows). Scale bars, 10  $\mu$ m.



**Extended Data Figure 8 | Interaction of RNF168 UDM1 with K63-ubiquitylated H1.** **a**, Immunoblot analysis of immobilized recombinant RNF168 UDM1 or UDM2 incubated with K63 linked di-ubiquitin (Ub<sub>2</sub>) in the presence of increasing salt concentrations (75 mM, 150 mM and 250 mM KCl, respectively). **b**, Binding of immobilized recombinant Strep-UDM2 or empty Strep-Tactin beads to indicated di-ubiquitin (Ub<sub>2</sub>) linkages was analysed by immunoblotting. **c**, Biotinylated peptides corresponding to the LRM1 and LRM2 motifs in human RNF168 were analysed for binding to recombinant H2A or H1.0 *in vitro* by Streptavidin pull-down followed by SDS-PAGE and

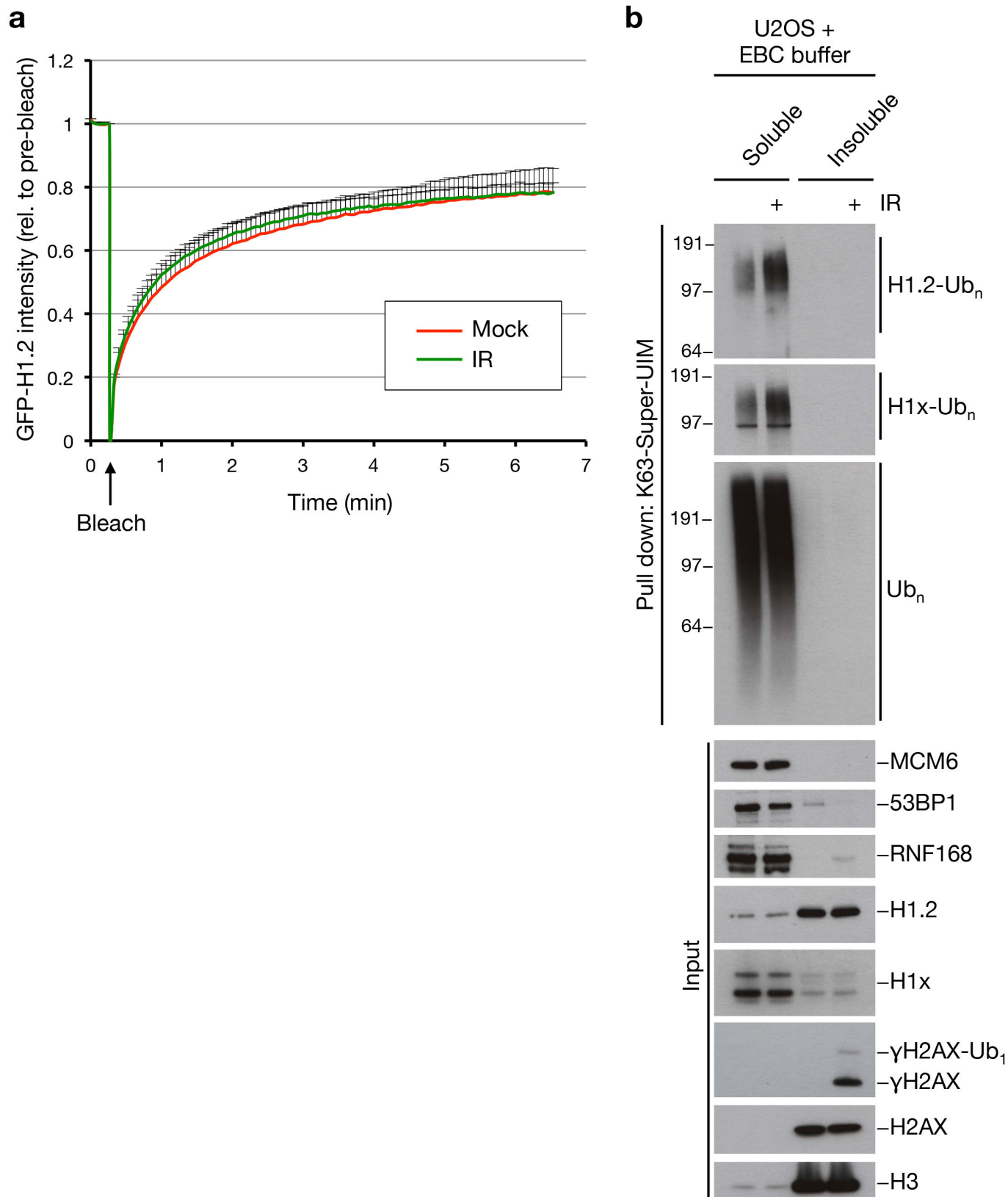
Colloidal Blue staining. The migration of molecular weight markers (kDa) is indicated on the left. **d**, Sequence of the UDM1 region in human RNF168, showing the location of the LRM1, UMI and MIU1 motifs. Acidic amino acids are highlighted in red. The sequence corresponding to the LRM1 peptide (c) and mutations introduced to generate UDM1 \*UMI and \*MIU1 (f) are indicated. **e**, **f**, Pull-down assays of Strep-tagged UDM1 and UDM2 constructs expressed in U2OS cells. **a-c**, **e**, **f**, Uncropped blots are shown in Supplementary Fig. 1.



**Extended Data Figure 9 | RNF168 UDM1 recognizes ubiquitylated forms of H1 but not core histones.** **a**, Pull-downs of Strep-tagged RNF168 UDM1 expressed in U2OS cells were immunoblotted (IB) with antibodies to indicated histones. **b**, Localization pattern of GFP-tagged UDM1 expressed in U2OS cells. Scale bar, 10  $\mu$ m. **c**, Venn diagram showing overlap between proteins

displaying increased K63-linked ubiquitylation after IR (SILAC ratio (IR/mock) > 1.5) and proteins showing potential interaction with overexpressed GFP-UDM1 (SILAC ratio (GFP-UDM1/mock) > 1.5). Only one protein, histone H1x, was common to both of these subsets of cellular proteins. **a**, Uncropped blots are shown in Supplementary Fig. 1.





**Extended Data Figure 10 | Impact of DSBs on H1 chromatin association.**  
**a**, FRAP analysis of U2OS cells stably expressing GFP-H1.2 and exposed or not to IR (10 Gy). Individual data points represent mean values from ten independent measurements and error bars represent twice the s.d. **b**, U2OS cells left untreated or exposed to IR were lysed in EBC buffer. Soluble and

resolubilized, EBC-insoluble fractions were incubated with recombinant K63-Super-UIM and washed thoroughly. Bound material and input fractions were analysed by immunoblotting with indicated antibodies. **b**, Uncropped blots are shown in Supplementary Fig. 1.

# The inner workings of the hydrazine synthase multiprotein complex

Andreas Dietl<sup>1</sup>, Christina Ferousi<sup>2</sup>, Wouter J. Maalcke<sup>2</sup>, Andreas Menzel<sup>3</sup>, Simon de Vries<sup>4,†</sup>, Jan T. Keltjens<sup>2</sup>, Mike S. M. Jetten<sup>2,4</sup>, Boran Kartal<sup>2,5</sup> & Thomas R. M. Barends<sup>1</sup>

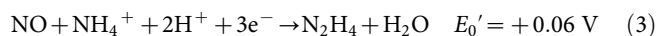
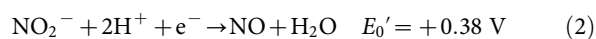
Anaerobic ammonium oxidation (anammox) has a major role in the Earth's nitrogen cycle<sup>1,2</sup> and is used in energy-efficient wastewater treatment<sup>3</sup>. This bacterial process combines nitrite and ammonium to form dinitrogen (N<sub>2</sub>) gas, and has been estimated to synthesize up to 50% of the dinitrogen gas emitted into our atmosphere from the oceans<sup>2</sup>. Strikingly, the anammox process relies on the highly unusual, extremely reactive intermediate hydrazine<sup>4</sup>, a compound also used as a rocket fuel because of its high reducing power. So far, the enzymatic mechanism by which hydrazine is synthesized is unknown. Here we report the 2.7 Å resolution crystal structure, as well as biophysical and spectroscopic studies, of a hydrazine synthase multiprotein complex isolated from the anammox organism *Kuenenia stuttgartiensis*. The structure shows an elongated dimer of heterotrimers, each of which has two unique c-type haem-containing active sites, as well as an interaction point for a redox partner. Furthermore, a system of tunnels connects these active sites. The crystal structure implies a two-step mechanism for hydrazine synthesis: a three-electron reduction of nitric oxide to hydroxylamine at the active site of the γ-subunit and its subsequent condensation with ammonia, yielding hydrazine in the active centre of the α-subunit. Our results provide the first, to our knowledge, detailed structural insight into the mechanism of biological hydrazine synthesis, which is of major significance for our understanding of the conversion of nitrogenous compounds in nature.

Most nitrogen on earth occurs as gaseous N<sub>2</sub> (nitrogen oxidation number 0). To make nitrogen available for biochemical reactions, the inert N<sub>2</sub> has to be converted to ammonia (oxidation number −III), which can then be assimilated to produce organic nitrogen compounds, or be oxidized to nitrite (oxidation number +III) or nitrate (+V). The reduction of nitrite in turn results in the regeneration of N<sub>2</sub>, thus closing the biological nitrogen cycle.

To produce N<sub>2</sub> from nitrite, a nitrogen–nitrogen bond must be formed by the addition of another nitrogen-containing molecule. At present, two biological processes are known that can achieve this. In denitrification, nitrite is first reduced to nitric oxide (NO, +II). Then, two molecules of NO are combined to produce nitrous oxide (N<sub>2</sub>O, +I), which is subsequently reduced to N<sub>2</sub>. The other process, anaerobic ammonium oxidation or anammox<sup>1,2</sup>, was discovered only relatively recently, and relies on the combination of two compounds with different nitrogen oxidation states, nitrite and ammonium, to generate N<sub>2</sub>.

Our current understanding of the anammox reaction (equation (1)) is based on genomic, physiological and biochemical studies on the anammox bacterium *K. stuttgartiensis*<sup>4,5</sup>. First, nitrite is reduced to nitric oxide (NO, equation (2)), which is then condensed with ammonium-derived ammonia (NH<sub>3</sub>) to yield hydrazine (N<sub>2</sub>H<sub>4</sub>, equation (3)). Hydrazine itself is a highly unusual metabolic intermediate, as it is extremely reactive and therefore toxic, and has a very low redox poten-

tial ( $E_0' = -750$  mV). In the final step in the anammox process, it is oxidized to N<sub>2</sub>, yielding four electrons (equation (4)) that replenish those needed for nitrite reduction and hydrazine synthesis and are used to establish a proton-motive force across the membrane of the anammox organelle, the anammoxosome, driving ATP synthesis (see ref. 6 for a review).



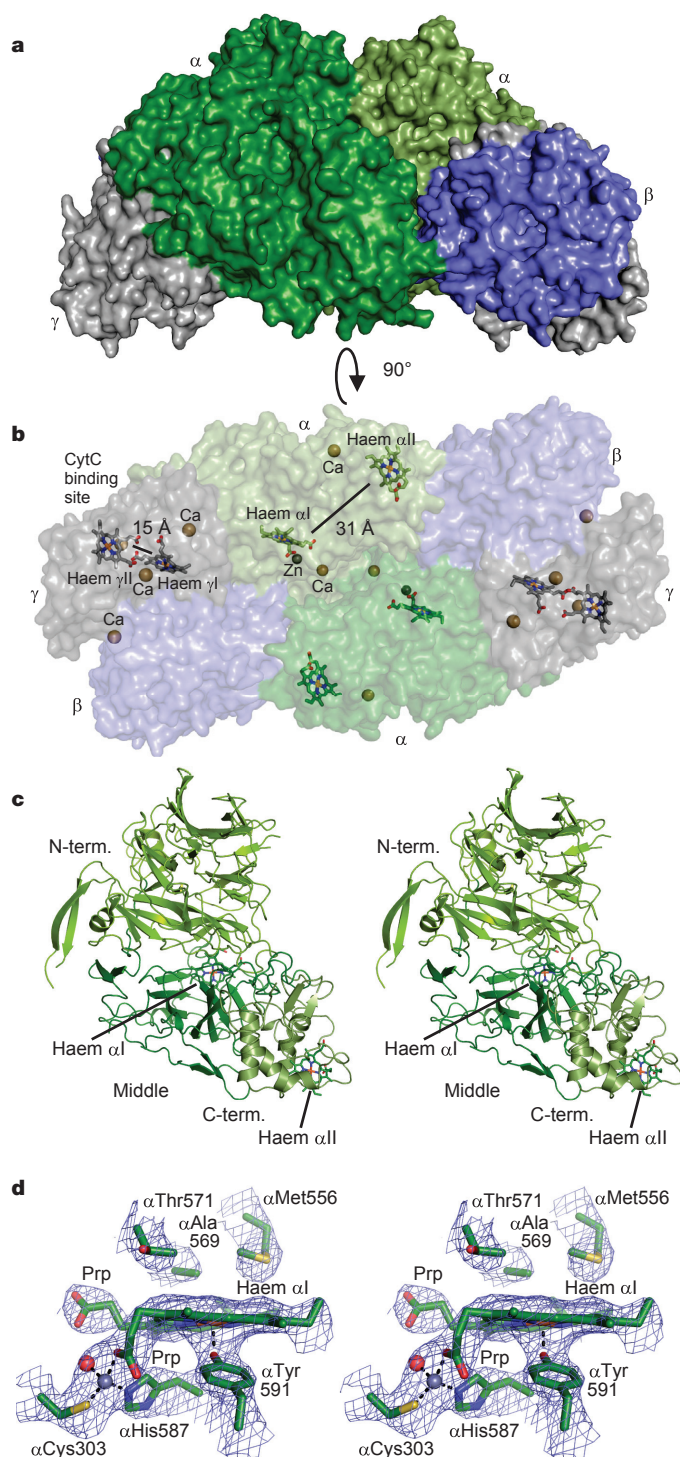
The enzyme producing hydrazine from NO and ammonium—hydrazine synthase (HZS)—is biochemically unique. A complex of three proteins, HZS-α, -β and -γ, encoded by the genes *kuste2861*, *-2859* and *-2860*, respectively, was put forward as the probable hydrazine synthase enzyme<sup>5</sup>. This complex was isolated from *K. stuttgartiensis* cells and shown to be catalytically active in a coupled assay with the octahaem c-type cytochrome *kustc1061* (ref. 7) to convert hydrazine into N<sub>2</sub> and return electrons to HZS<sup>4</sup>. Isolated HZS is a comparatively slow enzyme with an activity of 20 nmol h<sup>−1</sup> mg<sup>−1</sup> protein, about 1% of *in vivo* turnover. This striking loss of activity occurs immediately upon cell lysis and might be explained by the disruption of a tightly coupled multicomponent system, as well as by the use of bovine cytochrome *c* as an artificial electron carrier in the *in vitro* assay<sup>4</sup>.

Using a custom-designed crystal cooling method, we prepared well-diffracting crystals of the HZS-αβγ multienzyme complex from *K. stuttgartiensis* and determined its crystal structure at 2.7 Å resolution in the absence of substrates (Fig. 1a and Extended Data Table 1). The structure reveals a crescent-shaped dimer of heterotrimers with an (αβγ)<sub>2</sub> stoichiometry. The overall size and shape of the complex were confirmed by analytical ultracentrifugation and solution small-angle X-ray scattering (Supplementary Information and Extended Data Fig. 1). Each heterotrimer contains four haems and one zinc ion, as well as several calcium ions (Fig. 1b, Supplementary Information and Extended Data Table 2).

The α-subunit (Fig. 1c) consists of three domains: an N-terminal domain which includes a six-bladed β-propeller, a middle domain binding a pentacoordinated c-type haem (haem αI) and a C-terminal domain which harbours a bis-histidine-coordinated c-type haem (haem αII). The structure around haem αI (Fig. 1d) deviates substantially from a typical haem *c* site, as the canonical histidine of the haem *c* binding motif, αHis587, is rotated away from the haem iron, and coordinates a zinc ion. Instead, the hydroxyl group of αTyr591 serves as the proximal ligand to the haem iron, as in the active site of

<sup>1</sup>Department of Biomolecular Mechanisms, Max Planck Institute for Medical Research, 69120 Heidelberg, Germany. <sup>2</sup>Department of Microbiology, Institute for Water and Wetland Research, Radboud University Nijmegen, 6525 AJ Nijmegen, The Netherlands. <sup>3</sup>Swiss Light Source, Paul Scherrer Institute, 5232 Villigen, Switzerland. <sup>4</sup>Department of Biotechnology, Delft University of Technology, Delft, The Netherlands. <sup>5</sup>Department of Biochemistry and Microbiology, Laboratory of Microbiology, Gent University, Gent, Belgium.

†Deceased.



**Figure 1 | Crystal structure of HZS.** **a**, HZS complex structure;  $\alpha$ -subunits are coloured green,  $\beta$ -subunits are blue and  $\gamma$ -subunits are grey. **b**, Surface representation. The contact area between two heterotrimers ( $\sim 1,350 \text{ \AA}^2$ ) is made up of contributions from  $\alpha$ - and  $\beta$ -subunits only. Considerable solvent-filled space remains between the heterotrimers. Calcium ions are labelled Ca, zinc as Zn. Edge-to-edge distances between the haems within a subunit are indicated in ångströms. **c**, Stereofigure of the  $\alpha$ -subunit. The N-terminal domain (residues  $\alpha 28$ – $420$ ), middle domain ( $\alpha 421$ – $\alpha 670$ ) and C-terminal domain ( $\alpha 671$ – $808$ ) are indicated in different shades of green. The two haem groups are shown as sticks. **d**, Stereofigure of the haem  $\alpha I$  site, overlaid with the simulated annealing composite omit map, contoured at  $1.0\sigma$ . The zinc ion and its coordinating water are shown as grey and red spheres, respectively. The haem propionates are labelled Prp.

many catalases<sup>8</sup>. Importantly, this tyrosine is conserved in HZS- $\alpha$  sequences (Extended Data Fig. 2). The zinc bound to  $\alpha\text{His}587$  is further coordinated by one of the haem  $\alpha I$  propionate groups, as well as  $\alpha\text{Cys}303$  and probably a water molecule, in a structure reminiscent of the active sites of alcohol dehydrogenase and various metalloproteases<sup>9</sup>. The zinc ion could play a structural role, assisting in rotating  $\alpha\text{His}587$  away from the iron, allowing  $\alpha\text{Tyr}591$  to bind, or could directly modulate the chemistry of the haem group, with which it interacts via a propionate group.  $\alpha\text{Thr}571$ ,  $\alpha\text{Ala}569$  and  $\alpha\text{Met}556$  (which is partially oxidized, see Supplementary Information) are in close proximity to the distal side of haem  $\alpha I$ , which does not seem to coordinate a solvent molecule in the crystal structure. In contrast, haem  $\alpha II$  is bound by a canonical haem *c* binding motif and is coordinated by  $\alpha\text{His}772$  distally and  $\alpha\text{His}689$  proximally. The edge-to-edge distance<sup>10</sup> between haems  $\alpha I$  and  $\alpha II$  is  $31 \text{ \AA}$  (Fig. 1b), which is too long for single-step electron transfer between the haem groups of the  $\alpha$ -subunit. The edge-to-edge distances between the haem groups in the two different  $\alpha$ -subunits in the complex are larger than  $38 \text{ \AA}$ , which excludes electron transfer between the two  $\alpha$ -subunits on the timescale of catalysis.

The non-haem  $\beta$ -subunit (Fig. 2a) is a seven-bladed  $\beta$ -propeller with a short helical insertion in the sixth propeller blade. The outer strand of the C-terminal blade consists of the N terminus (residues  $\gamma 40$ – $52$ ) of the  $\gamma$ -subunit of the same heterotrimer. Notably, the HZS  $\beta$ - and  $\gamma$ -subunits are fused into a single polypeptide in the anammox bacteria *Scalindua profunda* and *Scalindua brodae* (ref. 11 and Extended Data Fig. 3).

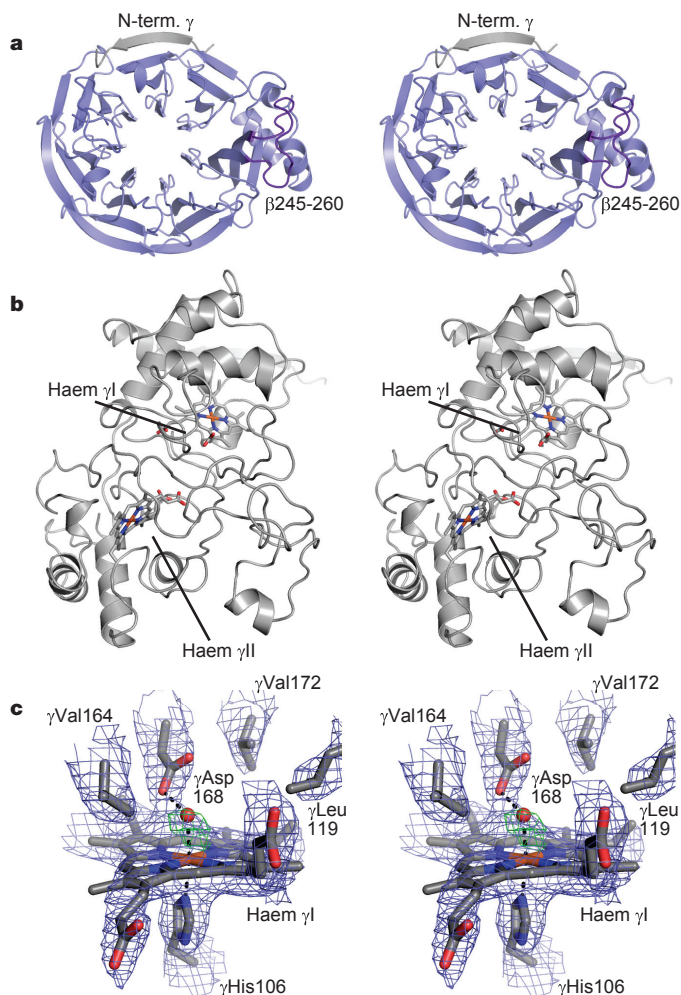
The structure of the  $\gamma$ -subunit (Fig. 2b) is reminiscent of the fold of the homologous dihaem cytochrome *c* peroxidases (CCPs)<sup>12–14</sup> and *Paracoccus denitrificans* methylamine utilization protein G (MauG)<sup>15</sup> and consists of two  $\alpha$ -helical lobes, each of which contains one *c*-type haem. Haem  $\gamma I$  in the N-terminal lobe (Fig. 2c) is coordinated proximally by  $\gamma\text{His}106$  and distally by a water molecule, and is covalently bound to  $\gamma\text{Cys}102$  and  $\gamma\text{Cys}105$  on a typical haem *c* binding motif. Intriguingly, the electron density maps clearly show a unique third covalent bond with the protein, between the  $C_1$  porphyrin methyl group and the  $S_\gamma$  sulfur atom of  $\gamma\text{Cys}165$  (Extended Data Fig. 4a), which possibly serves to modulate haem chemistry. At the distal side, the iron binds a water molecule, which is hydrogen bonded to  $\gamma\text{Asp}168$ . This conserved residue (Extended Data Fig. 3) is perfectly positioned to transfer protons to a ligand molecule coordinated to the haem. A structural superposition (Extended Data Fig. 4b) reveals that haem  $\gamma I$  is located at the position of the high-spin haem of the homologous *Nitrosomonas europaea* CCP<sup>13</sup> and *P. denitrificans* MauG<sup>15</sup>.

The bis-His-coordinated haem  $\gamma II$  in the C-terminal lobe is located at the equivalent position as the electron transfer haem in CCPs and MauG (Extended Data Fig. 4b), at an edge-to-edge distance of  $15 \text{ \AA}$  from haem  $\gamma I$  (Fig. 1b), which would allow direct electron transfer between the haems in the  $\gamma$ -subunit. In CCPs and MauG, a conserved Trp residue is believed to be involved in catalytic redox chemistry. In HZS- $\gamma$ , the position of this tryptophan is taken up by  $\gamma\text{His}144$ . The  $\gamma$ -subunit binds three calcium ions, one of them at the same position as the Ca-binding site in CCP that is essential for its activation. Moreover, haem  $\gamma II$  is located on the surface of the complex, exposed to the solvent, surrounded by a negatively charged patch, as in a cytochrome *c* binding site (Extended Data Fig. 5). Therefore, haem  $\gamma II$  probably functions in electron transfer.

Thus, it appears that the  $\alpha$ - and the  $\gamma$ -subunit each contain an active site (haems  $\alpha I$  and  $\gamma I$ ) and the  $\gamma$ -subunit contains an electron-transfer site (haem  $\gamma II$ ). Electron paramagnetic resonance (EPR) spectroscopy (Extended Data Fig. 6 and Supplementary Information) is consistent with a stoichiometry of two bis-His-coordinated haems and two haems for which a population of ligation states exist.

Intriguingly, our crystal structure revealed a tunnel connecting the haem  $\alpha I$  and  $\gamma I$  sites (Fig. 3a). This tunnel branches off towards the surface of the protein approximately halfway between the haem sites,

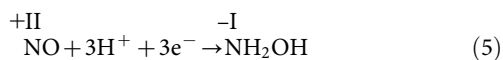




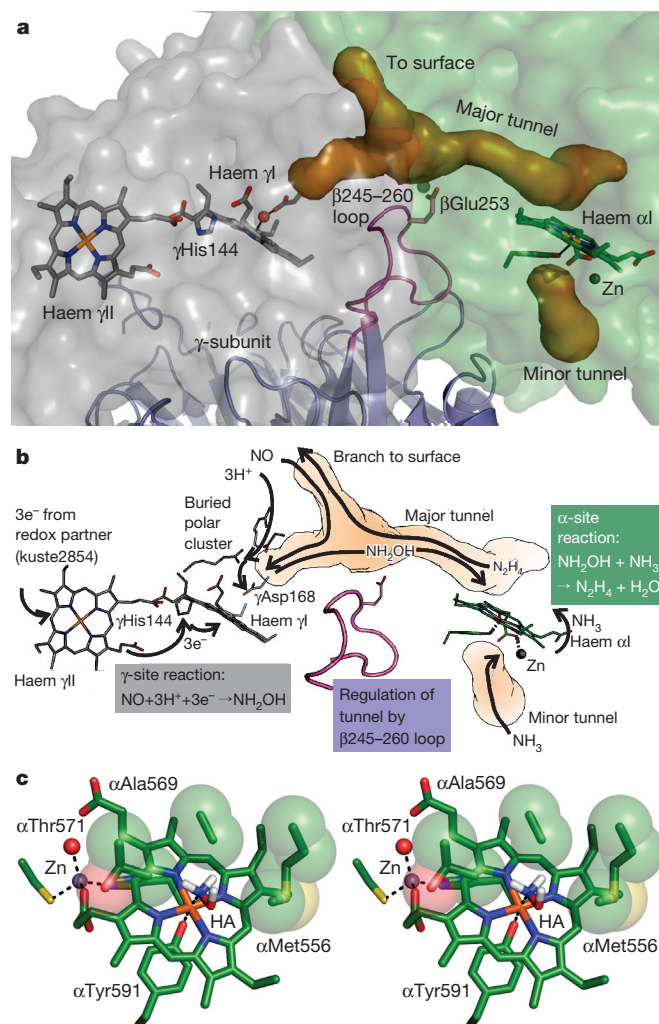
**Figure 2 | Structure of HZS- $\beta$  and HZS- $\gamma$ .** **a**, Structure of the  $\beta$ -subunit. The  $\beta$ 245–260 insertion is shown in purple. The N terminus of the  $\gamma$ -subunit, which engages in  $\beta$ -completion with the first blade of the  $\beta$ -propeller is shown in grey. **b**, Structure of the  $\gamma$ -subunit. **c**, Stereoview of haem  $\gamma$ I and its surroundings, overlaid with the simulated annealing composite omit map (blue,  $1.0\sigma$ ). The water molecule bound to the haem iron is shown as a red sphere. The green mesh is the difference electron density calculated before inclusion of the water molecule in the model ( $5.0\sigma$ ).

making them accessible to substrates from the solvent. Indeed, binding studies show that haem  $\alpha$ I is accessible to xenon (Extended Data Fig. 4c). Interestingly, in-between the  $\alpha$ - and  $\gamma$ -subunits, the tunnel is approached by a 15-amino-acid-long loop of the  $\beta$ -subunit ( $\beta$ 245–260), placing the conserved  $\beta$ Glu253, which binds a magnesium ion, into the tunnel.

These observations allow a mechanism for biological hydrazine synthesis to be proposed (Fig. 3b). The presence of two active sites, connected by a tunnel, strongly suggests a mechanism with two half-reactions. HZS combines NO (nitrogen oxidation number +II) and  $\text{NH}_4^+$  (N oxidation number –III). To reach the –II oxidation number of the nitrogen atoms in hydrazine, nitric oxide must be reduced. As proposed earlier<sup>6</sup>, this could happen in the  $\gamma$ -subunit, resulting in the production of hydroxylamine ( $\text{NH}_2\text{OH}$ ; nitrogen oxidation number –I) according to equation (5).



This three-electron reduction is consistent with the proposal that HZS obtains electrons from the trihaem cytochrome *c* kuste2854 (ref. 6). In this scheme, the electrons would enter HZS through haem  $\gamma$ II and be

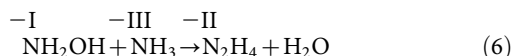


**Figure 3 | Proposed mechanism of biological hydrazine synthesis.** **a**, Tunnel between the active site haems (orange, major tunnel) with the branch to the protein surface. The  $\beta$ 245–260 loop is shown in purple, as well as  $\beta$ Glu253 which binds a magnesium ion (light-green sphere) and  $\gamma$ His144. A minor tunnel (lower right) leads to the zinc ion, and could allow ammonium to enter. **b**, Details of the proposed mechanism. NO travels to haem  $\gamma$ I through the tunnel (orange) via the branch leading to the surface. On the left, three electrons enter the complex at haem  $\gamma$ II and are conducted to haem  $\gamma$ I via  $\gamma$ His144. Together with three protons reaching haem  $\gamma$ I from the solvent via the buried polar cluster, the electrons reduce NO to  $\text{NH}_2\text{OH}$  (grey box).  $\text{NH}_2\text{OH}$  then diffuses through the tunnel, which is regulated by the  $\beta$ -subunit through the  $\beta$ 245–260 loop, and binds to haem  $\alpha$ I. There, it undergoes comproportionation with  $\text{NH}_3$  to yield hydrazine (green box). **c**, Stereoview, showing a model of hydroxylamine (HA) bound to haem  $\alpha$ I in a very hydrophobic environment.

transferred to the active site haem  $\gamma$ I, possibly via  $\gamma$ His144.  $\gamma$ Asp168 could assist in adding the protons. A cluster of buried, polar residues ( $\gamma$ Asp112,  $\gamma$ Arg143 and  $\gamma$ Arg167) is positioned between  $\gamma$ Asp168 and the surface of the complex and could serve to transfer protons to the active centre of the  $\gamma$ -subunit.

In the proposed mechanism, hydroxylamine then diffuses through the tunnel to the  $\alpha$ -subunit's active site. Given the position of the  $\beta$ 245–260 loop, the  $\beta$ -subunit could play a role in modulating transport through the tunnel. Hydroxylamine is isoelectronic with hydrogen peroxide, and is a competitive catalase inhibitor<sup>16</sup>. Thus, it would bind to the distal coordination site of the catalase-like haem  $\alpha$ I, which would polarize the N–O bond. As crystal soaking with  $\text{NH}_2\text{OH}$  was unsuccessful, we constructed a model of this complex

(Fig. 3c) which shows that hydroxylamine would be bound in a tight, very hydrophobic pocket, so that there is little electrostatic shielding of the partial positive charge on the nitrogen. Ammonia produced from ammonium (the predominant form at pH = 6.3 in the anammoxosome<sup>17</sup>) could then perform a nucleophilic attack on the nitrogen of hydroxylamine, yielding hydrazine through comproportionation (equation (6)).



Hydrazine could leave the enzyme via the tunnel branch leading to the surface.

Interestingly, the proposed scheme is analogous to the Raschig process used in industrial hydrazine synthesis. There, ammonia is oxidized to chloramine (NH<sub>2</sub>Cl, nitrogen oxidation number −I, like in hydroxylamine), which then undergoes comproportionation with another molecule of ammonia to yield hydrazine.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 1 April; accepted 21 August 2015.**

**Published online 19 October 2015.**

1. Devol, A. H. Denitrification, anammox, and N<sub>2</sub> production in marine sediments. *Ann. Rev. Mar. Sci.* **7**, 403–423 (2015).
2. Lam, P. & Kuypers, M. M. M. Microbial nitrogen cycling processes in oxygen minimum zones. *Ann. Rev. Mar. Sci.* **3**, 317–345 (2011).
3. Kartal, B., Kuenen, J. G. & van Loosdrecht, M. C. M. Sewage treatment with anammox. *Science* **328**, 702–703 (2010).
4. Kartal, B. *et al.* Molecular mechanism of anaerobic ammonium oxidation. *Nature* **479**, 127–130 (2011).
5. Strous, M. *et al.* Deciphering the evolution and metabolism of an anammox bacterium from a community genome. *Nature* **440**, 790–794 (2006).
6. Kartal, B. *et al.* How to make a living from anaerobic ammonium oxidation. *FEMS Microbiol. Rev.* **37**, 428–461 (2013).
7. Maalcke, W. J. *et al.* Structural basis of biological NO generation by octaheme oxidoreductases. *J. Biol. Chem.* **289**, 1228–1242 (2014).
8. Putnam, C. D., Arvai, A. S., Bourne, Y. & Tainer, J. A. Active and inhibited human catalase structures: ligand and NADPH binding and catalytic mechanism. *J. Mol. Biol.* **296**, 295–309 (2000).
9. Auld, D. S. Zinc coordination sphere in biochemical zinc sites. *Biometals* **14**, 271–313 (2001).
10. Moser, C. C., Chobot, S., Page, C. & Dutton, L. Distance metrics for heme protein electron tunneling. *Biochim. Biophys. Acta* **1777**, 1032–1037 (2008).
11. van de Vossenberg, J. *et al.* The metagenome of the marine anammox bacterium ‘*Candidatus Scalindua profunda*’ illustrates the versatility of this globally important nitrogen cycle bacterium. *Environ. Microbiol.* **15**, 1275–1289 (2013).
12. Fülöp, V., Ridout, C. J., Greenwood, C. & Hajdu, J. Crystal structure of the di-heme cytochrome c peroxidase from *Pseudomonas aeruginosa*. *Structure* **3**, 1225–1233 (1995).
13. Shimizu, H. *et al.* Crystal structure of *Nitrosomonas europaea* cytochrome c peroxidase and the structural basis for ligand switching in bacterial di-heme peroxidases. *Biochemistry* **40**, 13483–13490 (2001).
14. Echalié, A. *et al.* Redox-linked structural changes associated with the formation of a catalytically competent form of the di-heme cytochrome c peroxidase from *Pseudomonas aeruginosa*. *Biochemistry* **47**, 1947–1956 (2008).
15. Jensen, L. M. R., Sanishvili, R., Davidson, V. L. & Wilmot, C. M. In crystallo posttranslational modification within a MauG/Pre-methylamine dehydrogenase complex. *Science* **327**, 1392–1394 (2010).
16. Blaschko, H. The mechanism of catalase inhibitions. *Biochem. J.* **29**, 2303–2312 (1935).
17. van der Star, W. R. L. *et al.* An intracellular pH gradient in the anammox bacterium *Kuenenia stuttgartiensis* as evaluated by <sup>31</sup>P NMR. *Appl. Microbiol. Biotechnol.* **86**, 311–317 (2010).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We dedicate this work to Simon de Vries, who passed away unexpectedly shortly before the publication of this paper. The Dortmund-Heidelberg data collection team and the staff of beamline X10SA at the Swiss Light Source of the PSI in Villigen, Switzerland are acknowledged for their help and facilities. M. Gradl and M. Müller are thanked for assistance with MALDI- and ESI-TOF mass spectrometric analyses. We thank I. Schlichting, J. Reimann, M. Cryle, J. Reinstein and R. Shoeman for suggestions and C. Kieser (electronics workshop at MPIImF) for constructing the Peltier cooling controller used in post-crystallization treatment. T.R.M.B. thanks I. Schlichting for continuous support. B.K. and W.J.M. were supported by the Netherlands Organization for Scientific Research (VENI grant 863.11.003 and Darwin grant 142.16.1201, respectively). C.F. and M.S.M.J. are supported by the European Research Council (ERC232937) and by a Spinoza Prize awarded to M.S.M.J. Chimera is developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIGMS P41-GM103311). This work was supported by the Max Planck Society.

**Author Contributions** C.F. and W.J.M. isolated the Kuste2859-60-61 complex from *K. stuttgartiensis*. A.D. and T.R.M.B. performed X-ray crystallographic, SAXS and AUC analyses. A.M. performed SAXS measurements. C.F. performed ICP-MS analyses. W.J.M. and S.deV. performed EPR sample preparation and analysis. A.D. and T.R.M.B. wrote the paper with input from M.S.M.J., J.T.K., S.deV., C.F., W.M. and B.K.

**Author Information** The atomic coordinates and structure factors have been deposited in the Protein Data Bank, under accession codes 5C2V and 5C2W. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to T.R.M.B. (Thomas.Barends@mpiimf-heidelberg.mpg.de) or B.K. (kartal@science.ru.nl).

## METHODS

**Protein purification.** The kuste2859-2860-2861 hydrazine synthase (HZS) complex was purified from a planktonic *K. stuttgartiensis* culture as described previously<sup>4</sup>. Briefly, cell-free extracts prepared from a ~95% single-cell enrichment culture of *K. stuttgartiensis* were subjected to ultracentrifugation (180,000g; 4 °C; 1 h) to pellet cell membranes. HZS present in the bright-red supernatant was brought to homogeneity by a two-step column purification procedure consisting of subsequent Q Sepharose XL (GE Healthcare) and CHT Ceramic Hydroxyapatite (Bio-Rad, USA) column chromatography steps. UV-Vis spectra of as-isolated HZS showed a Soret absorption peak at 406 nm and a broad band in the 530 nm region, which are typical for fully oxidized (ferric) haem *c* proteins. Reduction of the protein under anoxic conditions using sodium dithionite resulted in a shift of the Soret maximum to 420 nm as well as haem  $\alpha$ - and  $\beta$ -bands at 553 nm and 523 nm, respectively. Protein concentrations used for ICP-MS and EPR measurements were determined using the Bradford assay (Bio-Rad, USA) with bovine serum albumin as standard.

**Analyses by MALDI-TOF and ESI-TOF mass spectrometry.** The subunits of the HZS complex were separated by 15% sodium dodecylsulfate polyacrylamide gel electrophoresis (SDS-PAGE). To identify the individual subunits of the HZS complex, Coomassie-stained SDS gel slices were digested with trypsin or chymotrypsin, followed by reduction with DTT and alkylation with iodoacetamide. The resulting peptides were purified and concentrated using a Millipore ZipTip C-18 column, spotted onto solid targets with  $\alpha$ -cyanocinnamic acid, and analysed by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) on an Axima TOF<sup>2</sup> Performance mass spectrometer (Shimadzu Biotech, Duisburg, Germany). Signal peptide cleavage sites were predicted using the SignalP 3.0 Server<sup>18</sup> applying Hidden-Markov models for Gram-negative bacteria. Liquid HZS samples were analysed by electrospray ionization time-of-flight mass spectrometry (ESI-TOF MS) on a maXis spectrometer (Bruker Daltonics) under denaturing conditions after diluting in 50% (v/v) acetonitrile/0.1% formic acid and separation by reversed-phase high-performance liquid chromatography (RP-HPLC) using a Discovery BIO Wide Pore C5 column (20  $\times$  2.1 mm, 5  $\mu$ m particle size, Supelco) at a flow rate of 50  $\mu$ l min<sup>-1</sup>.

**Metal analysis by inductively-coupled plasma mass spectrometry (ICP-MS).** Metals were analysed by ICP-MS on a Series I ICP MS (Thermo Scientific, Breda, the Netherlands). Height point calibration was performed with a dilution series of (multi-) element standards (1,000 p.p.b. in 1% nitric acid; Merck, Darmstadt, Germany) using the PlasmaLab software (Thermo Scientific, Breda, the Netherlands). To determine the metal content of HZS, 70–300  $\mu$ l of purified HZS (26 mg protein per ml) was washed with water using a Vivaspin 500 filter (Sartorius, Göttingen, Germany), destructed with 10% nitric acid at 90 °C for 90 min and diluted to 10 ml with water.

**Analytical ultracentrifugation (AUC).** Protein samples were concentrated in 25 mM Hepes/KOH, pH 7.5, 25 mM KCl to  $A_{280}^{1\text{cm}} \approx 0.3$  and  $A_{406}^{1\text{cm}} \approx 0.45$ , corresponding to 0.3 mg ml<sup>-1</sup>, as determined using the Bradford protein assay from Bio-Rad. Sedimentation velocity analytical ultracentrifugation was performed in a Beckman ProteomeLab XL-I (Beckmann Coulter, Krefeld, Germany) analytical ultracentrifuge equipped with an An60Ti rotor at 30,000 r.p.m. and 20 °C in a two-sector cell with a 1.2 cm optical path length. Absorption scan data were collected at 280 nm and 406 nm and evaluated using SEDFIT<sup>19</sup>.

**Protein crystallization and crystal treatment.** Hydrazine synthase was concentrated to 45 mg ml<sup>-1</sup> in 25 mM HEPES/KOH pH 7.5, 25 mM KCl by ultrafiltration, divided into 50- $\mu$ l aliquots, frozen in liquid nitrogen and stored at -80 °C. Prior to crystallization, the protein stock was diluted to 30 mg ml<sup>-1</sup> with 25 mM HEPES/KOH pH 7.5, 25 mM KCl. Crystallization was performed in 1  $\mu$ l + 1  $\mu$ l sitting drop vapour-diffusion setups at 8 °C, equilibrating against 500  $\mu$ l 36% (v/v) 1,4-dioxane. Dark red, rhombohedral crystals with dimensions up to 400  $\times$  400  $\times$  100  $\mu$ m grew within three days. The addition of 1 mM 5-amino-2,4,6-triiodoisophthalic acid adjusted to pH 7 with ethanolamine to the precipitant solution in the drops yielded more crystals and accelerated crystal growth. Using PEG 400 or other conventional cryoprotectants, diffraction of these crystals suffered from diffuse scattering, limiting resolution to approx. 4 Å and precluding SAD phasing. Successful cryoprotection was carried out by soaking the crystals for 10–30 s in 4 M betaine (*N,N,N*-trimethylglycine) in 50% (v/v) methanol at 8 °C, before flash-cooling in liquid nitrogen. These crystals showed sharp Bragg spots, were used for phasing and to build the initial model. However, as the crystals dissolved in the soaking solution at 8 °C, crystals were slowly cooled to -20 °C on a custom-designed Peltier-cooled microscope stage, which will be described in detail elsewhere. After incubation in the soaking solution at this temperature for up to 30 min, crystals were flash-cooled in liquid propane at a temperature of approximately 150 K. These crystals diffracted up to 2.7 Å resolution. Xenon treatment was performed in a -20 °C room by transferring the crystals cooled to -20 °C into a Xe-pressure cell (Xcell, Oxford Cryosystems Ltd, Long

Hanborough, UK) and incubating for 5 min at -20 °C and a xenon pressure of 20 bar before freezing in liquid propane.

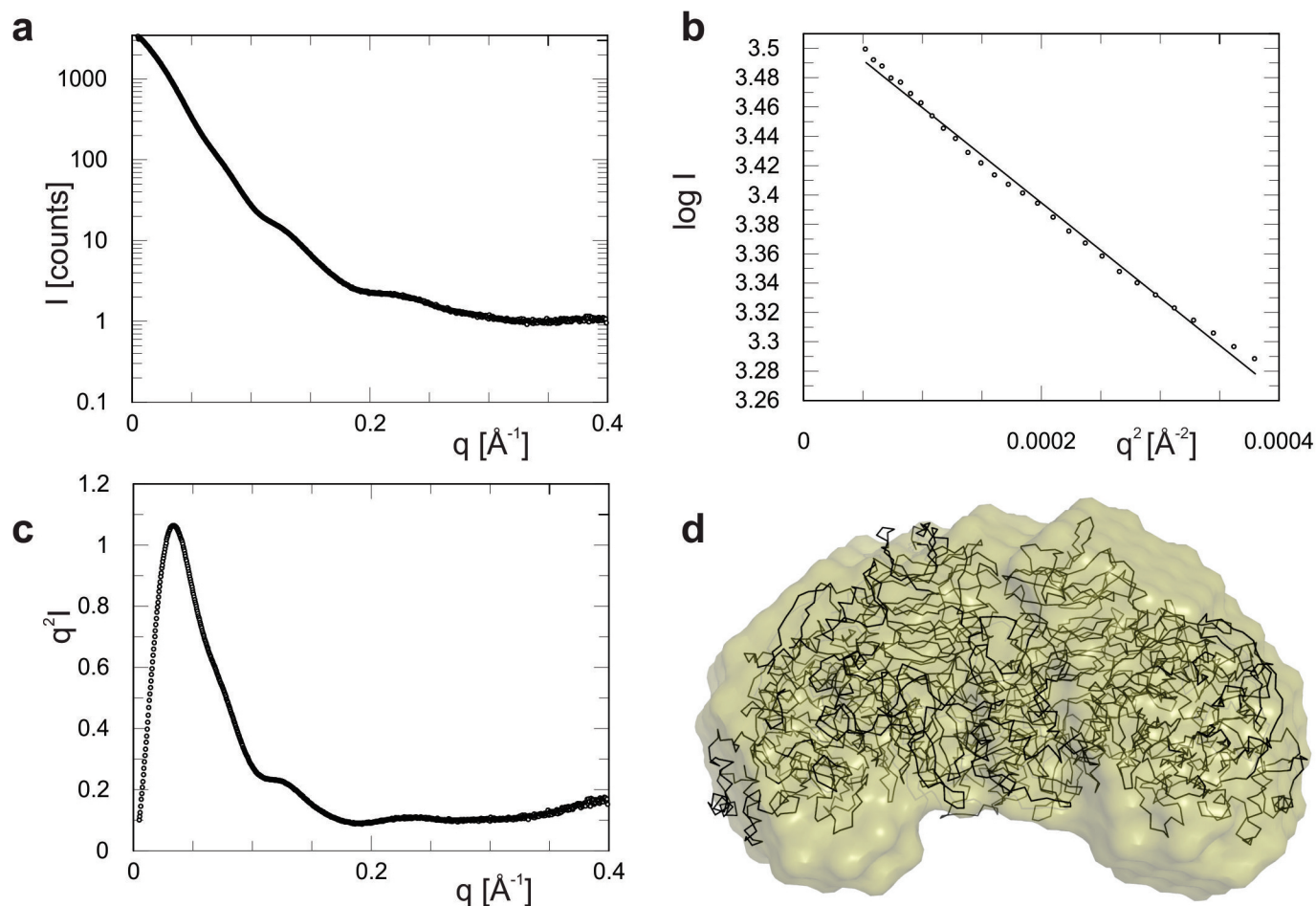
**X-ray data collection, structure solution and analysis.** Diffraction data were collected at beam line X10SA of the Swiss Light Source (Paul Scherrer Institute, Villigen, Switzerland) at 100 K and processed with XDS<sup>20</sup>. A highly redundant single-wavelength anomalous dispersion (SAD) data set at a resolution of 3.7 Å was collected just above the iron K-edge at a wavelength of 1.735 Å (see Extended Data Table 1) which was used for phase determination with AutoSHARP<sup>21</sup>. SHELXD<sup>22</sup> identified 5 heavy atom sites (CC(E) = 0.24), which were used by SHARP for phasing, resulting in a figure-of-merit of 0.22. Density modification with SOLOMON<sup>23</sup> resulted in a readily interpretable map, into which the structures of all three subunits could be built using Coot<sup>24</sup>. Phase extension using a data set of 3.1 Å collected at 0.9763 Å wavelength was carried out with DM<sup>25</sup>. Further refinement against a 2.7 Å data set collected at 1.0000 Å wavelength using PHENIX<sup>26</sup> and REFMAC<sup>27</sup> resulted in a model with good geometry and R-factors (96.4% of residues in favoured regions of the Ramachandran plot, 0.07% Ramachandran outliers, see Extended Data Table 1) and revealed that two loop regions in the  $\alpha$ -subunit ( $\alpha$ 175–177 and  $\alpha$ 643–650) were no longer ordered, despite the increase in overall resolution. In order to confirm the identity of the metal sites, data sets above and below the K-edges of iron, copper and zinc were collected (see Supplementary Information and Extended Data Table 2). All other data-processing procedures were performed with programs of the CCP4 suite<sup>28</sup>. Tunnels were identified using MOLE 2.0 (ref. 29) using standard parameter settings starting from  $\alpha$ Thr571,  $\alpha$ Tyr591 and  $\gamma$ Asp168. Structural figures were prepared using Pymol (Schrödinger). The model of the hydroxylamine complex was prepared by manual docking in COOT<sup>24</sup>, using an iron–nitrogen bond length between haem  $\alpha$ I and hydroxylamine as observed in crystal structures of catalase–hydroxylamine complexes.

**Small-angle X-ray scattering (SAXS).** Hydrazine synthase was concentrated to 45 mg ml<sup>-1</sup> in 25 mM HEPES/KOH pH 7.5, 25 mM KCl. SAX data were measured in 1-mm diameter quartz capillaries at the X12SA beam line (cSAXS) of the Swiss Light Source (Paul Scherrer Institute, Villigen, Switzerland) at 283 K. The X-ray photon energy was 12.4 keV, and 200 measurements of 0.5 s each were recorded over 10 positions along the length of the capillary, which was mounted at a detector distance of 2.138 m. Background measurements with the buffer only were taken using the identical capillaries, positions and measurement protocol. Data were used to a maximum momentum transfer of 0.4 Å<sup>-1</sup>. Data analysis and three-dimensional reconstruction were performed using the GNOM<sup>30</sup> and GASBOR<sup>31</sup> programs from the ATSAS suite.

**EPR spectroscopy.** EPR spectroscopy was performed with a Varian E-9 spectrometer operating at X-band (microwave frequency 9.188 GHz; modulation amplitude, 1.0 mT) equipped with a home-made He-flow cryostat at 12 K. HZS samples were used as isolated at a concentration of 205  $\mu$ M, filled into quartz tubes and shock frozen in liquid nitrogen before the measurements. Samples in the presence of 200  $\mu$ M NH<sub>2</sub>OH or 200  $\mu$ M NO plus 200  $\mu$ M NH<sub>4</sub><sup>+</sup> were prepared and analysed in the same way.

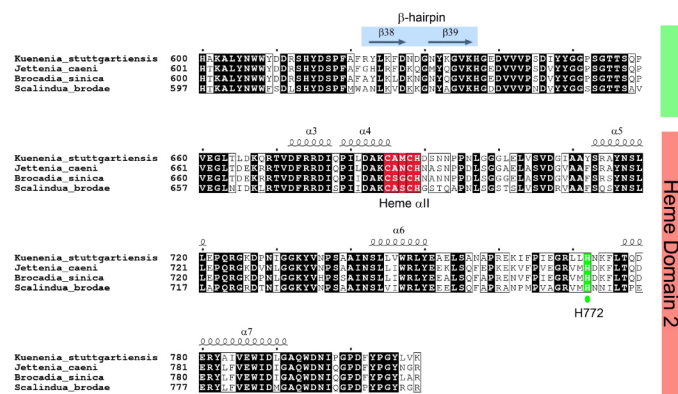
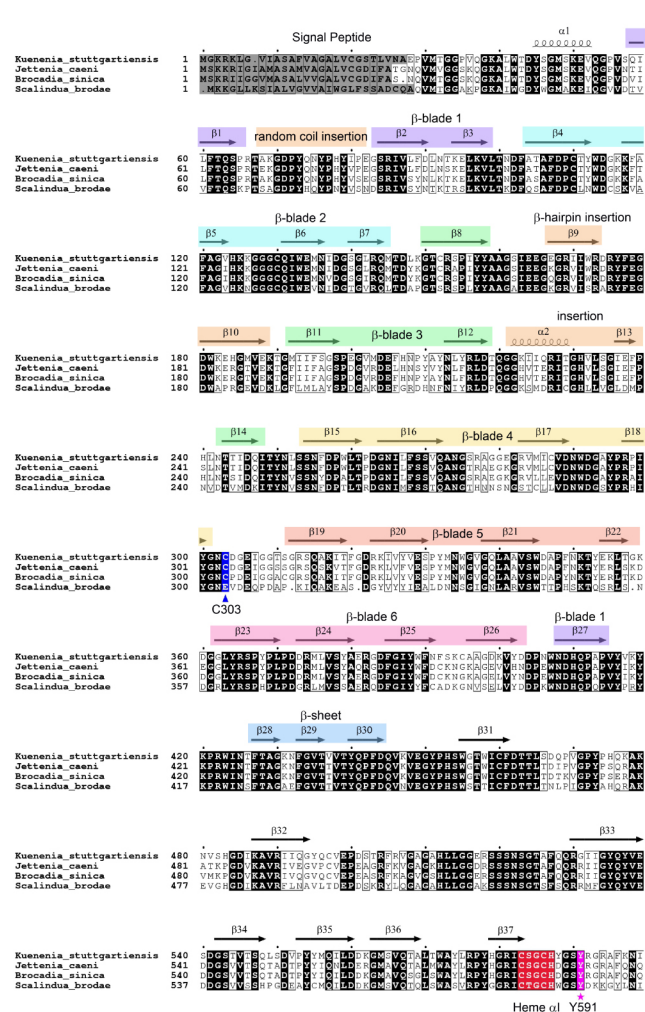
- Bendtsen, J. D., Nielsen, H., von Heijne, G. & Brunak, S. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* **340**, 783–795 (2004).
- Brown, P. H. & Schuck, P. Macromolecular size-and-shape distributions by sedimentation velocity analytical ultracentrifugation. *Biophys. J.* **90**, 4651–4661 (2006).
- Kabsch, W. XDS. *Acta Crystallogr. D* **66**, 125–132 (2010).
- Vonrhein, C., Blanc, E., Roversi, P. & Brice, G. Automated structure solution with autoSHARP. *Methods Mol. Biol.* **364**, 215–230 (2007).
- Schneider, T. R. & Sheldrick, G. M. Substructure solution with SHELXD. *Acta Crystallogr. D* **58**, 1772–1779 (2002).
- Abrahams, J. P. & Leslie, A. G. W. Methods used in the structure determination of bovine mitochondrial F<sub>1</sub> ATPase. *Acta Crystallogr. D* **52**, 30–42 (1996).
- Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
- Cowan, K. D. & Zhang, K. Y. J. Density modification for macromolecular phase improvement. *Prog. Biophys. Mol. Biol.* **72**, 245–270 (1999).
- Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
- Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53**, 240–255 (1997).
- Collaborative Computational Project, Number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
- Petřek, M., Kosinova, P., Koca, J. & Otyepka, M. MOLE: a Voronoi diagram-based explorer of molecular channels, pores, and tunnels. *Structure* **15**, 1357–1363 (2007).
- Svergun, D. I. Determination of the regularization parameter in indirect-transform methods using perceptual criteria. *J. Appl. Cryst.* **25**, 495–503 (1992).
- Svergun, D. I., Petoukhov, M. V. & Koch, M. H. J. Determination of domain structure of proteins from X-ray solution scattering. *Biophys. J.* **80**, 2946–2953 (2001).
- Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).





**Extended Data Figure 1 | Small-angle X-ray Scattering (SAXS) results.** **a**, Semilogarithmic plot of scattered intensity  $I$  versus  $q$ , which was defined as  $q = (4\pi \sin \theta)/\lambda$ . The curve is an average over 200 measurements. Features are observed up to  $q = 0.4 \text{ \AA}^{-1}$ . **b**, Guinier plot (plot of  $\log I$  versus  $q^2$ ) showing

that the protein is not aggregated. **c**, Kratky plot (plot of  $q^2 I$  versus  $q$ ) showing that the protein is folded. **d**, Average of 18 (out of 20) dummy-atom reconstructions (beige) overlaid on the crystal structure (black).



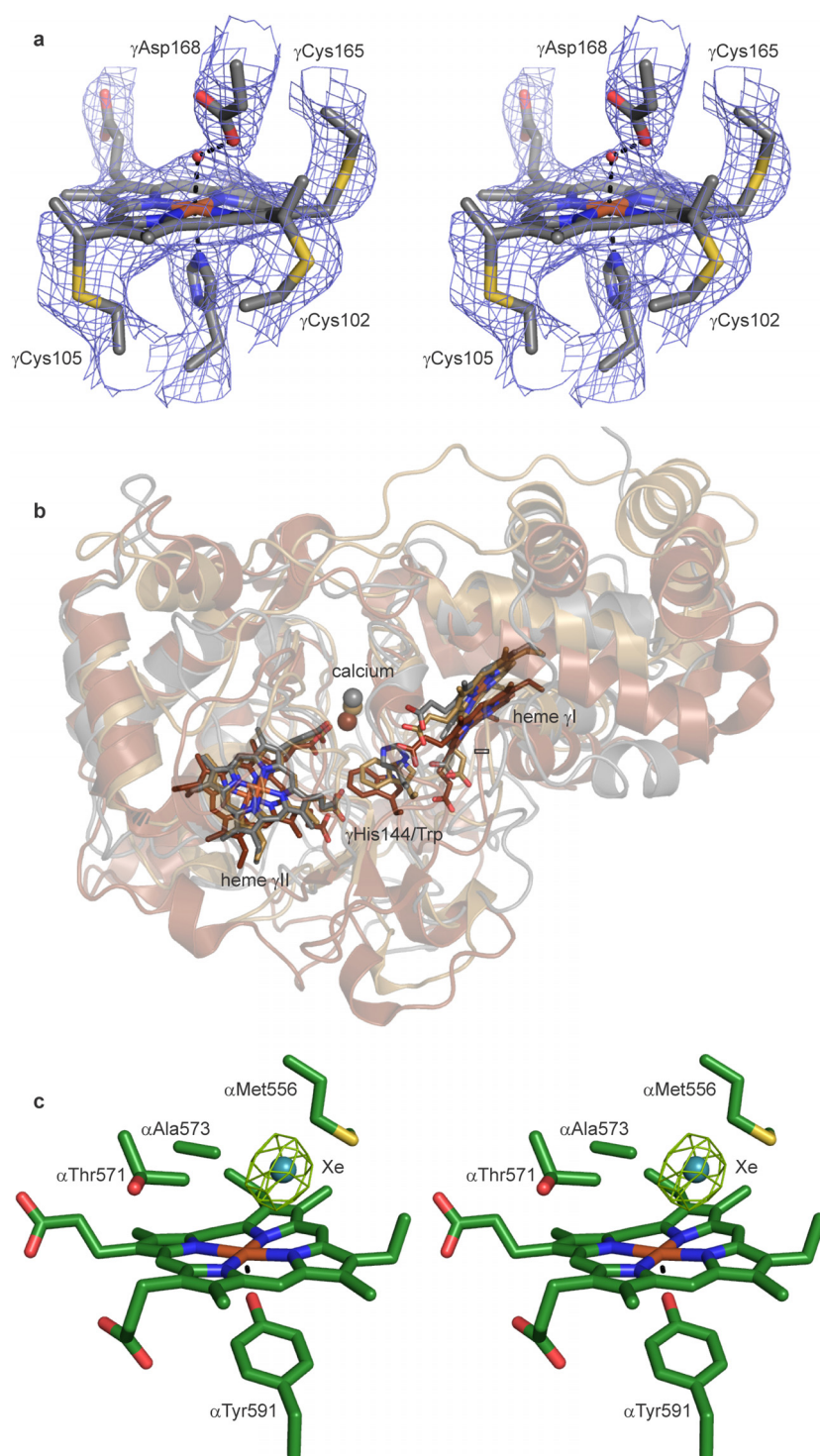
**Extended Data Figure 2 | HZS-α sequences.** The HZS-α sequences from *Kuenenia stuttgartiensis* (kuste2861, gi 91200564), *Jettenia caeni* (Planctomycete KSU-1, ksu1d0439, tr A9ZRZ5), *Brocadia sinica* JPN1 (brosiA2676, gi 762182098) and *Scalindua brodae* (scabro01598, gi 726045835, re-confirmed by Sanger sequencing) were aligned in ClustalW and secondary structure elements were manually assigned based on the structure of *Kuenenia* HZS-α. Kuste2861 shares 81% sequence identity with its *Jettenia* and

*Brocadia* orthologues and 61% with *Scalindua*. Fully conserved peptide sequences are marked black. The predicted signal peptides are highlighted in grey. The following residues are marked (numbering according to kuste2861): Cys303 coordinating  $Zn^{2+}$  (blue triangle), Tyr591 coordinating haem αI (pink asterisk), distal His772 of haem αII (green circle). The c-type haem binding motifs are highlighted in red. The figure was prepared using ESPrict.



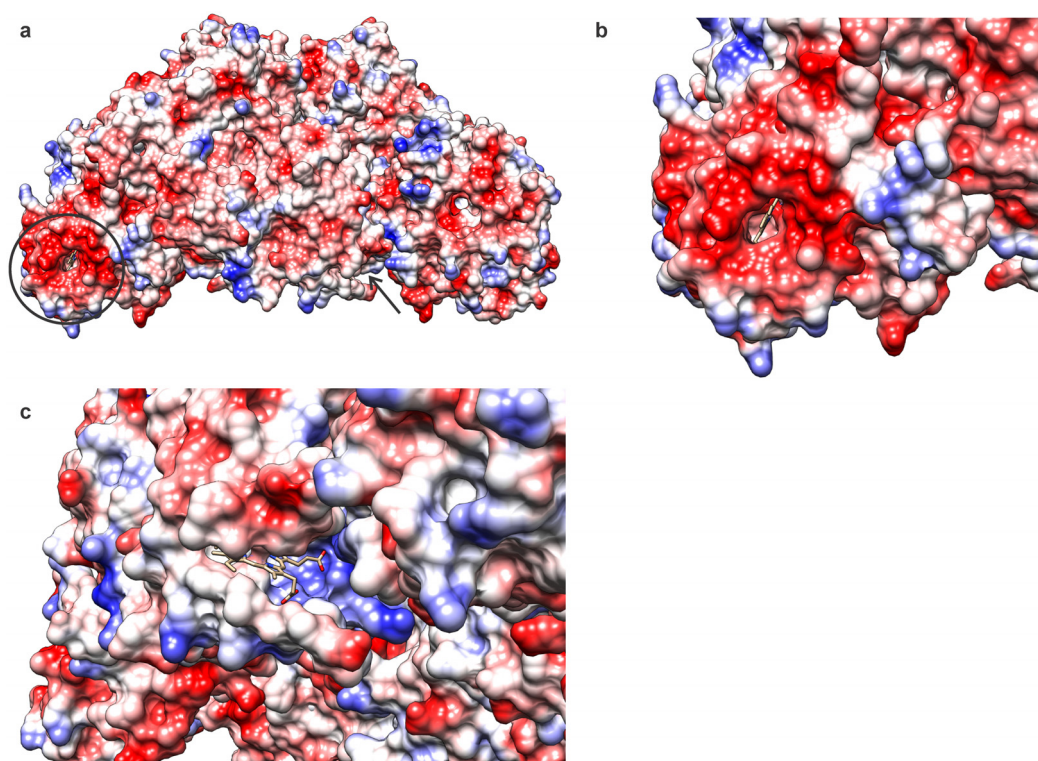
sequence indicate the numbering in kuste2860). The Kuste2859–60 fusion shares 83% sequence identity with its *J. caeni* and *B. sinica* orthologues and 72% with *S. brodae*. Fully conserved peptide sequences are marked in black. The predicted signal peptides of the  $\beta$ -subunits are highlighted in grey. The following residues are marked (numbering according to kuste2859 and kuste2860): Glu253 in HZS- $\beta$  (pink triangle), Cys165 covalently bound to haem  $\gamma$ I, Asp168 near the haem  $\gamma$ I catalytic site (blue triangle) and the distal His332 of haem  $\gamma$ II (green circle). The *c*-type haem binding motifs are highlighted in red. The figure was prepared using ESPript. The predicted signal peptides of the  $\gamma$ -subunits not included in the alignment are: kuste2860: MAREMRLGGKERMKTGVVKGILVAALGVVGLISAGGVYA\_GQP...; ksuId0440: MRNGMIKIGLVAALGIAGVVTAGEIMA\_GTP...; brosiA2675: MKSSLKIGLIAALGIAGVMTTGELMA\_GTP.





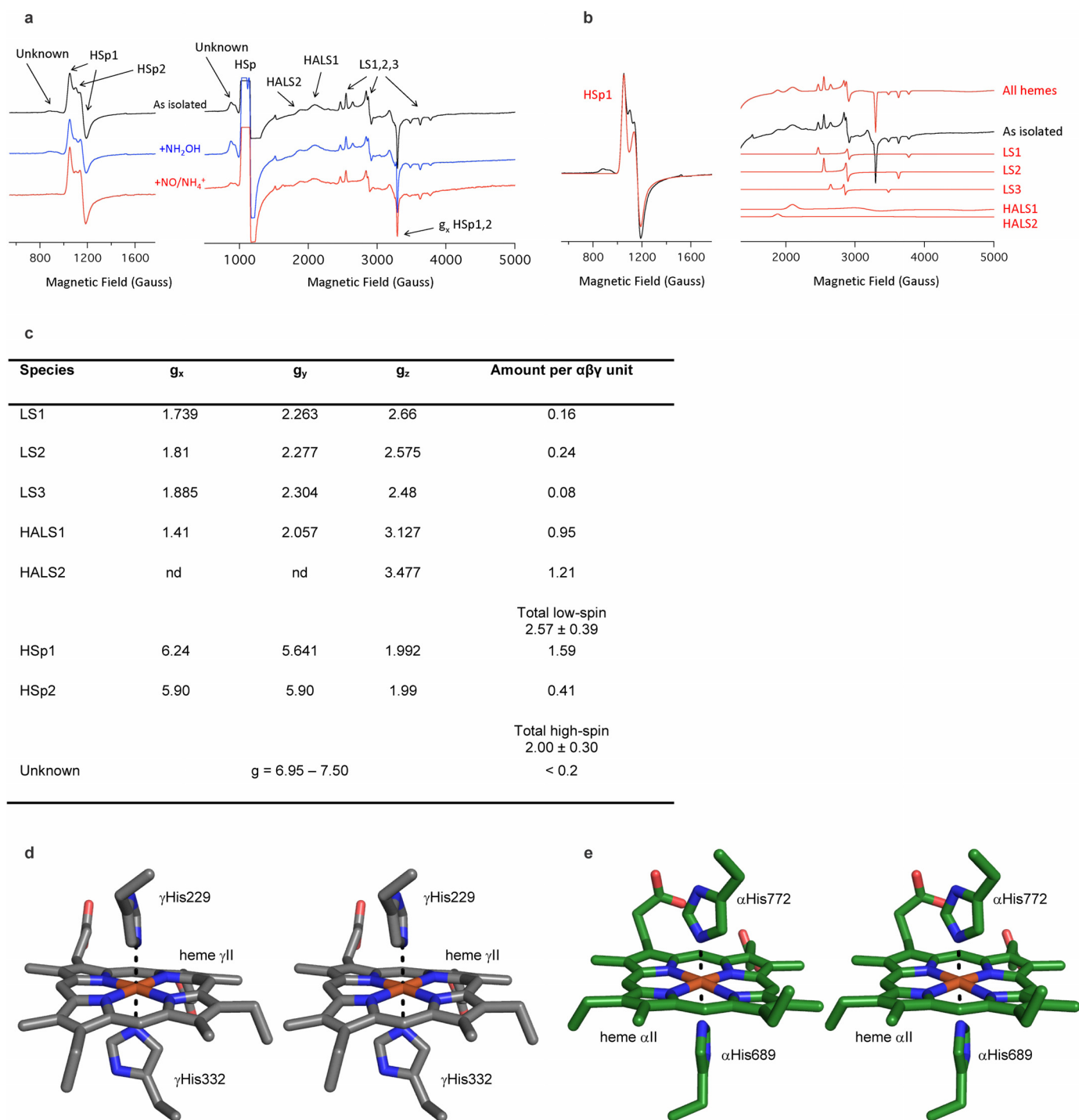
**Extended Data Figure 4 | Details of HZS structure.** **a**, Covalent attachment of haem  $\gamma$ I via three cysteine sulfur atoms. The simulated annealing  $2mF_o - DF_c$  composite omit map is shown contoured at  $1\sigma$ , overlaid on the final, refined structure.  $\gamma$ Cys102 and  $\gamma$ Cys105 are part of the canonical CXXCH motif (grey cartoon). In addition, there is a covalent bond between the S $\gamma$  atom of  $\gamma$ Cys165 and the C $_1$  porphyrin methyl group of haem  $\gamma$ I. **b**, Overlay of HZS  $\gamma$  (grey) with *N. europaea* CCP (PDB entry 1IQC, light brown) and *P. denitrificans* MauG (PDB entry 3L4M, dark brown). The positions of

haems  $\gamma$ I and  $\gamma$ II correspond to those of the haems in CCP and MauG (sticks), as does the position of a calcium ion (spheres). The conserved tryptophan residue proposed to be involved in redox catalysis in MauG and CCP corresponds to His144 in HZS- $\gamma$  (sticks). **c**, Xenon binding shows that haem  $\alpha$ I is accessible from the solvent. The Xe atom is shown as a sphere. Green mesh:  $mF_o - DF_c$  map calculated before inclusion of Xe in the model, ( $10\sigma$ ).  $\alpha$ Met556 has assumed a new conformation.



**Extended Data Figure 5 | Electrostatic surface properties of the HZS complex.** Haem moieties are shown as sticks. **a**, Overview of the whole HZS structure. The bis-His-coordinated haem  $\gamma$ II is indicated with a black circle. Haem  $\alpha$ II is obscured in this view but its position is indicated by a black arrow. **b**, Magnified view of the electrostatic properties of the surface

surrounding haem  $\gamma$ II. A prominent negatively charged patch surrounds the haem as in cytochrome *c* binding sites. **c**, Magnified view of the vacuum electrostatic properties of the surface surrounding haem  $\alpha$ II. No significant differences with the rest of the protein surface are observed. Figure prepared using UCSF Chimera<sup>32</sup>.



**Extended Data Figure 6 | EPR spectroscopy of HZS.** **a**, EPR spectra of HZS as isolated (black traces) and after addition of  $200 \mu\text{M}$   $\text{NH}_2\text{OH}$  (blue traces) or  $200 \mu\text{M}$   $\text{NO}$  plus  $200 \mu\text{M}$   $\text{NH}_4^+$  (red traces). The left panel shows the low magnetic field region highlighting the high-spin haem  $g_x$  and  $g_y$  resonances. The right panel shows the complete magnetic field scan where the intensity of the high-spin haem signals has run off-scale. Arrows indicate the positions of the various species that are listed in Extended Data Fig. 6c. The signal at 1540 Gauss is due to a small amount ( $<0.2\%$  per  $\alpha\beta\gamma$  unit) of adventitious iron. **b**, Simulation of the EPR spectra of HZS as isolated using the  $g$  values listed in Extended Data Fig. 6c. The difference between the simulation of HSp1 and the experimental spectrum defines the signal of HSp2 and its  $g$  value and suggests an amount of 0.41 per  $\alpha\beta\gamma$  unit (see panel c). **c**, HZS haem content

per  $\alpha\beta\gamma$  unit determined by EPR. The total haem content determined by EPR was  $0.92 \pm 0.15$  of the optically determined amount. nd, not detectable; LS, low-spin; HALS, highly anisotropic low-spin; HSp, rhombic high-spin peak. **d**, Stereofigure of the coordination of haem  $\gamma\text{II}$  by  $\gamma\text{His229}$  and  $\gamma\text{His332}$ . The perpendicular orientation of the histidine imidazole rings, both oriented towards haem *meso* atoms, is consistent with the  $g$ -values for HALS2. **e**, Stereofigure of the coordination of haem  $\alpha\text{II}$  by  $\alpha\text{His689}$  and  $\alpha\text{His772}$ . The orientation of the histidine imidazole groups, one ( $\alpha\text{His772}$ ) oriented towards a haem nitrogen atom and the other ( $\alpha\text{His689}$ ) towards a haem *meso* atom is consistent with the  $g$  values for HALS1 (see Supplementary Information).



Extended Data Table 1 | Data collection and refinement statistics

	SAD	Initial Model	Native structure (pdb 5C2V)	Xenon complex (pdb 5C2W)
<b>Data collection<sup>*</sup></b>				
Space group	<i>R</i> 32	<i>R</i> 32	<i>R</i> 32	<i>R</i> 32
Cell dimensions				
<i>a</i> , <i>b</i> , <i>c</i> (Å)	461.8, 461.8, 145.1	464.0, 464.0, 145.0	464.5, 464.5, 145.8	464.1, 464.1, 145.1
$\alpha$ , $\beta$ , $\gamma$ (°)	90, 90, 120	90, 90, 120	90, 90, 120	90, 90, 120
Resolution (Å)	40-3.4 (3.5-3.4) <sup>†</sup>	40-3.1(3.2-3.1)	40-2.7 (2.8-2.7)	48.5-3.2 (3.3-3.2)
<i>R</i> <sub>merge</sub>	0.103 (0.530)	0.100 (0.498)	0.096 (0.738)	0.138 (0.641)
<i>I</i> / $\sigma$ <i>I</i>	17.8 (3.2)	16.4 (3.7)	18.5 (3.5)	16.6 (4.6)
Completeness (%)	99.9 (100)	99.8 (99.8)	99.8 (100)	100.0 (100.0)
Redundancy	11.7 (8.1)	5.7 (5.6)	8.8 (9.1)	10.6 (10.8)
<b>Refinement</b>				
Resolution (Å)			40-2.7	48.5-3.2
No. reflections			162,788	97,821
<i>R</i> <sub>work</sub> / <i>R</i> <sub>free</sub>			0.235 / 0.271	0.231 / 0.267
No. atoms				
Protein			22,420	22,420
Ligand/ion			344 (8 heme)	344 (8 heme)
			18 (12 Ca, 2 Zn, 2 Mg, 2 Cl)	18 (12 Ca, 2 Zn, 2 Mg, 2 Cl)
			56 (7 betaines)	56 (7 betaines), 4 Xe
Water			500	498
B-factors (Å <sup>2</sup> )				
Protein			56.9	71.0
Ligand/ion			58.3	61.4
Water			53.4	57.4
R.m.s deviations				
Bond lengths (Å)			0.009	0.009
Bond angles (°)			1.2	1.3

\*Each data set was collected from a single crystal.

†Highest resolution shell is shown in parentheses.

Extended Data Table 2 | Identification of metals in hydrazine synthase

a

Data set*	Above Fe-edge	Below Fe-edge	Above Cu-edge
Space group	<i>R</i> 32	<i>R</i> 32	<i>R</i> 32
Unit cell dimensions			
<i>a</i> , <i>b</i> , <i>c</i> (Å)	467.2, 467.2, 146.0	465.3, 465.3, 145.5	465.3, 465.3, 145.5
$\alpha$ , $\beta$ , $\gamma$ (°)	90, 90, 120	90, 90, 120	90, 90, 120
Wavelength (Å)	1.73400	1.74600	1.37800
Resolution range (Å) †	30.0–3.80 (3.9–3.8)	30.0–3.50 (3.6–3.5)	30.0–3.20 (3.3–3.2)
Reflections measured	1,165,022 (88,940)	1,469,568 (107,411)	1,995,467 (172,227)
Reflections unique	116,983 (8,703)	147,985 (12,027)	193,705 (17,117)
Completeness (%)	99.9 (99.8)	99.9 (100)	100 (100)
Redundancy <i>N</i>	10.0 (10.2)	9.9 (8.9)	10.3 (10.1)
<i>I</i> / $\sigma$ <i>I</i>	14.4 (4.7)	18.6 (5.2)	22.2 (6.5)
<i>R</i> <sub>merge</sub> (%)	15.5 (59.7)	12.6 (50.1)	9.5 (43.6)

Data set	Below Cu-edge	Above Zn-edge	Below Zn-edge
Space group	<i>R</i> 32	<i>R</i> 32	<i>R</i> 32
Unit cell dimensions			
<i>a</i> , <i>b</i> , <i>c</i> (Å)	466.0, 466.0, 146.4	465.0, 465.0, 144.6	465.9, 465.9, 145.0
$\alpha$ , $\beta$ , $\gamma$ (°)	90, 90, 120	90, 90, 120	90, 90, 120
Wavelength (Å)	1.38500	1.28100	1.29433
Resolution range (Å) †	30.0–3.40 (3.5–3.4)	40.0–3.50 (3.6–3.5)	40.0–3.80 (3.9–3.8)
Reflections measured	1,669,421 (143,426)	1,465,592 (111,737)	1,134,042 (79,076)
Reflections unique	162,603 (13,535)	147,007 (11,917)	115,562 (8,699)
Completeness (%)	100 (100)	99.9 (100)	99.9 (100)
Redundancy <i>N</i>	10.3 (10.6)	10.0 (9.4)	9.8 (9.1)
<i>I</i> / $\sigma$ <i>I</i>	17.6 (5.6)	15.4 (3.8)	16.2 (4.6)
<i>R</i> <sub>merge</sub> (%)	13.1 (55.4)	13.3 (68.3)	13.2 (64.1)

b

Element/position wrt. absorption edge	Fe/below	Fe/above	Cu/below	Cu/above	Zn/below	Zn/above
Wavelength (Å)	1.746	1.734	1.385	1.378	1.29433	1.281
Energy (eV)	7101	7150	8952	8997	9579	9679
Energy difference from K-edge (eV)	-11	38	-27	18	-80	20
<b>HZS <math>\alpha</math></b>						
Ca $\alpha$ I	8.1/1.0	6.5/1.0	6.7/1.0	8.8/1.0	6.0/1.0	4.7/1.0
Ca $\alpha$ II	8.7/1.0	4.2/0.6	5.6/0.8	7.7/0.8	4.7/0.8	3.4/0.7
Zn	8.5/1.0	5.7/0.9	4.2/0.6	7.2/0.8	<b>4.5/0.8</b>	<b>24.3/5.2</b>
Fe Heme $\alpha$ I	<b>4.4/0.5</b>	<b>20.3/3.1</b>	20.5/3.1	27.4/3.1	17.4/2.9	16.3/3.5
Fe Heme $\alpha$ II	<b>3.8/0.5</b>	<b>18.8/2.9</b>	18.5/2.8	23.5/2.7	15.3/2.6	12.6/2.7
<b>HZS <math>\beta</math></b>						
Ca $\beta$ I	8.9/1.1	6.6/1.0	5.8/0.9	7.9/0.9	4.5/0.8	4.3/0.9
<b>HZS <math>\gamma</math></b>						
Ca $\gamma$ I	9.9/1.2	6.0/0.90	8.4/1.3	11.2/1.3	6.5/1.1	4.6/1.0
Ca $\gamma$ II	8.5/1.0	8.7/1.3	6.8/1.0	8.1/0.9	5.5/0.9	4.6/1.0
Ca $\gamma$ III	12.2/1.5	8.3/1.3	6.8/1.0	11.2/1.3	7.1/1.2	4.6/1.0
Fe Heme $\gamma$ I	<b>5.9/0.7</b>	<b>17.5/2.7</b>	20.4/3.0	27.0/3.1	14.0/2.3	13.8/2.9
Fe Heme $\gamma$ II	<b>5.5/0.7</b>	<b>16.2/2.5</b>	17.6/2.6	23.0/2.6	13.3/2.2	12.9/2.7

**a**, Data collection statistics for the anomalous diffraction data sets used to identify metal sites in the HZS crystal structure, calculated while considering Friedel mates as individual reflections. \*Each data set was collected from a single crystal, †Highest resolution shell is shown in parentheses. **b**, Heights of peaks in anomalous difference density maps used to identify metal ions. The first number is the observed peak height in  $\sigma$ , the second is the peak height normalized by the height of the anomalous peak at calcium  $\alpha$ I for each data set. Those sites that show a significant difference in normalized peak height below and above an absorption edge are shown in grey. The data confirm the identity of the zinc ion bound to haem  $\alpha$ I.

## Corrigendum: Whole-genome characterization of chemoresistant ovarian cancer

Ann-Marie Patch, Elizabeth L. Christie, Dariush Etemadmoghadam, Dale W. Garsed, Joshy George, Sian Fereday, Katia Nones, Prue Cowin, Kathryn Alsop, Peter J. Bailey, Karin S. Kassahn, Felicity Newell, Michael C. J. Quinn, Stephen Kazakoff, Kelly Quek, Charlotte Wilhelm-Benartzi, Ed Curry, Huei San Leong, The Australian Ovarian Cancer Study Group, Anne Hamilton, Linda Mileschkin, George Au-Yeung, Catherine Kennedy, Jillian Hung, Yoke-Eng Chiew, Paul Harnett, Michael Friedlander, Michael Quinn, Jan Pyman, Stephen Cordner, Patricia O'Brien, Jodie Leditschke, Greg Young, Kate Strachan, Paul Waring, Walid Azar, Chris Mitchell, Nadia Traficante, Joy Hendley, Heather Thorne, Mark Shackleton, David K. Miller, Gisela Mir Arnau, Richard W. Tothill, Timothy P. Holloway, Timothy Semple, Ivon Harliwong, Craig Nourse, Ehsan Nourbakhsh, Suzanne Manning, Senel Idrisoglu, Timothy J. C. Bruxner, Angelika N. Christ, Barsha Poudel, Oliver Holmes, Matthew Anderson, Conrad Leonard, Andrew Lonie, Nathan Hall, Scott Wood, Darrin F. Taylor, Qinying Xu, J. Lynn Fink, Nick Waddell, Ronny Drapkin, Euan Stronach, Hani Gabra, Robert Brown, Andrea Jewell, Shivashankar H. Nagaraj, Emma Markham, Peter J. Wilson, Jason Ellul, Orla McNally, Maria A. Doyle, Ravikiran Vedururu, Collin Stewart, Ernst Lengyel, John V. Pearson, Nicola Waddell, Anna deFazio, Sean M. Grimmond & David D. L. Bowtell

*Nature* **521**, 489–494 (2015); doi:10.1038/nature14410

In this Article, the affiliations of authors Michael Quinn and Orla McNally should read “<sup>22</sup>Department of Obstetrics and Gynaecology, The University of Melbourne, and The Royal Women’s Hospital, Parkville, Victoria 3052, Australia”. Their affiliations have been corrected in the HTML and PDF versions online.



## CORRIGENDUM

doi:10.1038/nature15720

# Corrigendum: Improving survival by exploiting tumour dependence on stabilized mutant p53 for treatment

E. M. Alexandrova, A. R. Yallowitz, D. Li, S. Xu, R. Schulz, D. A. Proia, G. Lozano, M. Dobbstein & U. M. Moll

*Nature* **523**, 352–356 (2015); doi:10.1038/nature14430

In this Letter on page 353, the words ‘substrate Hsp90’ should have been included in this sentence as follows: “Likewise, histone deacetylase inhibitors, including FDA-approved SAHA, are promising anti-cancer drugs whose actions involve hyperacetylation of histone and select non-histone targets including HDAC6 substrate Hsp90, thus indirectly inhibiting Hsp90 (ref. 21)”. This has been corrected in the online versions.

## ERRATUM

doi:10.1038/nature15718

### **Erratum: Arithmetic and local circuitry underlying dopamine prediction errors**

Neir Eshel, Michael Bukwich, Vinod Rao, Vivian Hemmelder, Ju Tian & Naoshige Uchida

*Nature* **525**, 243–246 (2015); doi:10.1038/nature14855

In this Letter, the *x*-axis labels of Fig. 3b and d were incorrect; for both panels, the labels should read: “Time from odour onset (s)”. Figure 3 has been corrected in the HTML and PDF versions online.

## ERRATUM

doi:10.1038/nature15719

# Erratum: A positional Toll receptor code directs convergent extension in *Drosophila*

Adam C. Paré, Athea Vichas, Christopher T. Fincher,  
Zachary Mirman, Dene L. Farrell, Avantika Mainieri  
& Jennifer A. Zallen

*Nature* **515**, 523–527 (2014); doi:10.1038/nature13953

In this Article, two errors were introduced into the figures during the production process. In Fig. 3i, the *x* axis should alternate 'WT, 2, WT, 2' rather than 'WT, 6,8, WT, 6,8', and in Fig. 2g, the 6,8 bar should have one asterisk, rather than two. These errors have been corrected in the online versions of the paper.



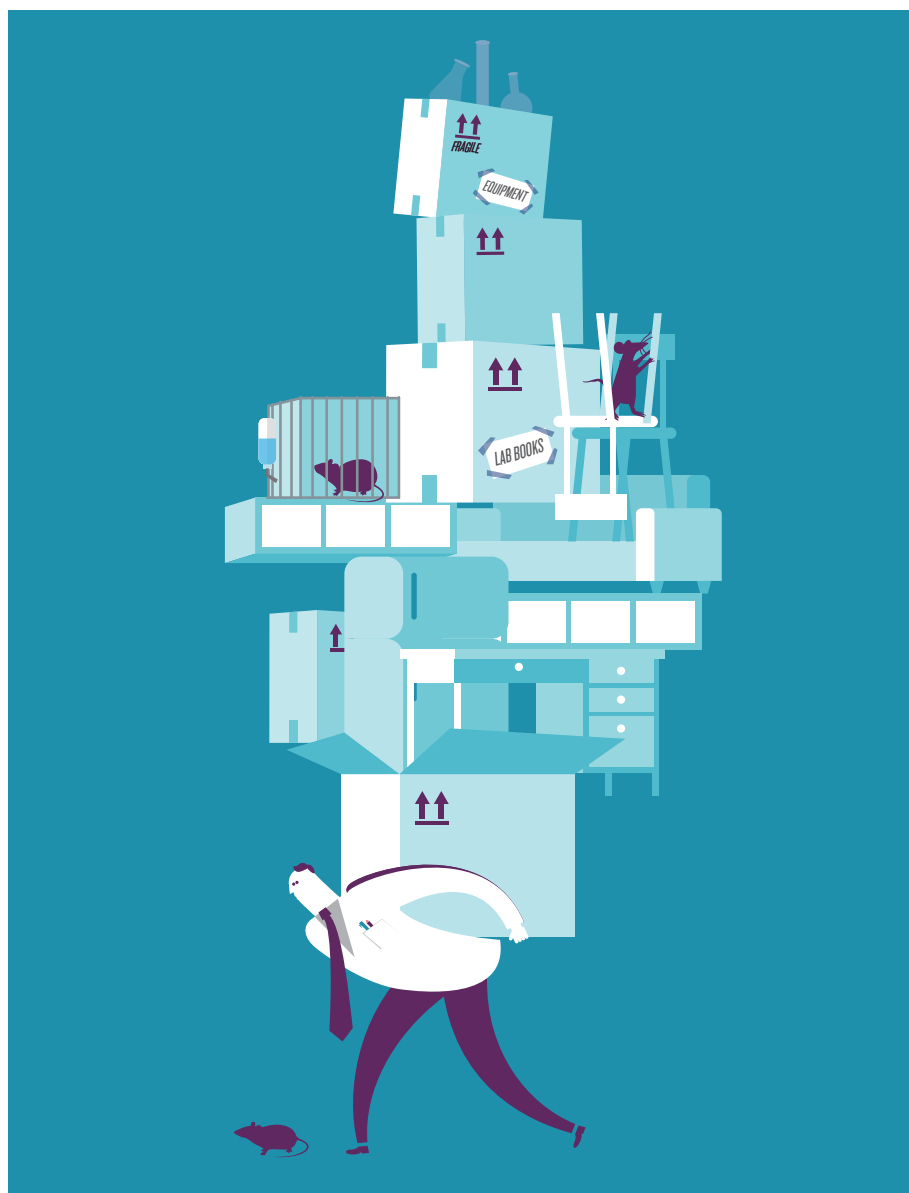
# CAREERS

**CARBON CONTROL** How a careful academic analysed Volkswagen's emissions **p.401**

**EQUAL OPPORTUNITIES** Women, science and a leaky pipeline [go.nature.com/kgewfo](http://go.nature.com/kgewfo)

**NATUREJOBS** For the latest career listings and advice [www.naturejobs.com](http://www.naturejobs.com)

ADAPTED FROM ON AKINDO/GETTY



## LAB TRANSITIONS

# The bumpy road to relocation

*Faced with a need to move lab, scientists should consider as early as possible how to effect a smooth transition.*

BY PAUL SMAGLIK

Two years into his PhD programme in immunology, Sudarshan Anand learned that his adviser was leaving the Mayo Clinic in Rochester, Minnesota, for Johns Hopkins University in Baltimore, Maryland. So where did that leave Anand? If he stayed at Mayo, he would need to find another mentor and probably another PhD project. But if he followed his adviser to Baltimore, he would have to rebuild his support system. "It is easier to make friends in the first year of grad school," he says. "Joining a new programme in year three, you need to be more proactive and make friends outside of the lab."

Anand followed his instincts — and his mentor — to Maryland. By doing so, he kept a research project and adviser whom he liked, but he was delayed by about a semester owing to the need to retake coursework. He also had to recruit a new thesis committee. "The faculty didn't know me. I just dropped in and said, 'Hey, I would like you to be on my committee.' That was a little tricky." The experience helped Anand to prepare for two subsequent relocations: one to the University of California, San Diego, for a postdoc in 2007, and another in 2013 to take a tenure-track position at Oregon Health & Science University in Portland, where he still works today.

Science is a mobile enterprise, and researchers at any stage of their careers could suddenly face the prospect of packing up and moving to a different institution, nation or even continent. The process is rarely easy, even for those who don't have the headache of moving an entire laboratory. For graduate students, a move may mean repeating coursework. For postdocs, it could mean losing access to painstakingly collected data, animal models or reagents, or sacrificing time to create back-up animal models or cell cultures. Senior scientists might need to recruit and train a new lab team. All those factors have a role in a scientist's decision on whether to move and, if so, how.

For Anand, things worked out well: his team at Mayo was close to publishing a paper when the time came to move, which helped to give him firm scientific footing. "It did wonders for my confidence," Anand says. His remaining time at Johns Hopkins felt more like a postdoc, he says, because he had time to do experiments that weren't essential to his dissertation. He passed his qualifying exams with minimal stress and few regrets about his circuitous ►

► educational path. “You learn a lot about yourself by how you handle curveballs,” he says.

But it isn't always such smooth sailing. Researchers who are faced with moving lab — whether to follow a mentor, because of a calamity or to snag a fellowship or research post — need to identify and maintain what they have already established before they consider recreating that situation in a new environment (see ‘Five simple steps for a fluid transition’).

## A CAN OF WORMS

Developmental biologist Phil Newmark left his first postdoc at the University of Barcelona in Spain to continue his work at the Carnegie Institution for Science in Baltimore. Naturally, he took his research collection of planarian flatworms (*Schmidtea mediterranea*) with him. But within two years, his entire worm colony had died off as a result of sudden problems with the in-house water-purification system — right as the team was making some important technical breakthroughs.

These were not just any worms. The Spanish species differs from the North American one in that it reproduces asexually, has a smaller genome and isn't easily procured. “One cannot simply buy these animals from a supplier,” Newmark says. “We could have lost years.” Panicked, he and his postdoc adviser flew back to Spain to make a pilgrimage to the broken fountain from which he had gathered his first batch. They were relieved to find the same type of planarian still dwelling in the standing water.

The problem recurred a few years later — again as a result of abrupt changes in local water quality — when Newmark moved to the

University of Illinois at Urbana–Champaign. This time, however, Newmark had a safety net of back-up worms, which gave him time to develop defined culture conditions, using ultrapure water as the starting point. That was fortuitous, because by then workers had repaired the fountain in Spain. It is no longer a habitat for planaria.

Biomolecular engineer W T. Godbey also had to adapt quickly to unforeseen disaster. In 2005, Hurricane Katrina destroyed his lab, forcing him to move temporarily from Tulane University in New Orleans, Louisiana, to Rice University in Houston, Texas. One of his graduate students, Xiujuan Zhang, decided to follow him — but she had to find him first. This was not an easy task in an age before ubiquitous mobile phones, and Tulane's e-mail server was down because of the flood. She eventually tracked him down through his mother.

“The first few weeks were the scariest part, because no one knew where anyone was — home, on vacation, dead,” Godbey recalls. Once the flood waters had receded, he visited his lab at Tulane to see how much damage the hurricane had caused. His plasmid samples had been wrapped in garbage bags, boxed and shoved into a hot, mouldy freezer in a sweltering laboratory — yet fortunately, none had seriously degraded. Back at Rice, he and Zhang grew the DNA segments into a

larger library, then sequenced the plasmids to ensure that none had been compromised. None had, but Godbey knows now that back-up supplies — and disaster planning — are essential.

## COUNTRY CONUNDRUMS

Even the most careful plans can be thrown out of whack when complications arise, especially for senior scientists. Molecular biologist Josh Brickman began a relocation in 2011 from the University of Edinburgh, UK, to the University of Copenhagen, where he had accepted a group-leader post at the then-newly created Danish Stem Cell Center. He had to transport multiple types of animal, recreate several mouse lines, resettle half-a-dozen lab staff and, for a time, supervise labs in two countries to fulfil dual grant commitments. “It was an experience,” he says.

Different animal-housing conventions at the two facilities intensified the stress inherent in moving more than 100 frogs and 6 lines of transgenic mice. The Edinburgh facility had housed the mice in open-topped cages, which risked exposing them to pathogens. The animal facility in Copenhagen, by contrast, had closed cages that were considered pathogen-free — which meant that the relocated mice could not be placed directly into them. Instead, the animals were mated, and their embryos were removed and transferred to surrogate mothers that had been raised in the Copenhagen pathogen-free lab. “It took much longer than we had expected,” Brickman says. “And we had problems doing mouse experiments during the process.”

Transporting the embryonic-stem-cell lines — which according to Brickman's estimates

*“The first few weeks were the scariest because no one one knew where anyone was — home, on vacation, dead.”*

## SMOOTH MOVES

### *Five simple steps for a fluid transition*

There is no one-size-fits-all approach to moving lab, but those who have successfully navigated the ordeal can offer useful advice.

#### ● **Create back-up plans for resources.**

Phil Newmark's planaria worms survived a move from Barcelona, Spain, to Baltimore, Maryland — until a change in the lab's water-purification system killed the lot. Now a developmental biologist at the University of Illinois at Urbana–Champaign, Newmark confirmed for himself that when moving it is crucial to consider resources — in his case, both his live specimens and water quality at his new site. “After our die-off, I made a big effort to make multiple clonal lines and establish long-term colonies,” he says.

#### ● **Make sure that your new lab is ready.**

Construction delays at biomolecular engineer Josh Brickman's new lab at the University of Copenhagen meant that his equipment — including animal models — had to be housed

in temporary facilities. That created an extra step in an already complicated move from the University of Edinburgh, UK. “If I were faced with doing it again, I would certainly move only once I knew the facility was finished,” says Brickman.

● **Keep lab staff in the loop.** Everyone should be informed as early as possible that you are considering a move, no matter your career stage. Talking openly about plans and considering how they might affect lab members and the lab's work is important for a smooth transition, whether or not the group is going, too. “They may have situations you're not aware of,” says W T. Godbey, an engineer at Tulane University in New Orleans, Louisiana. He had to move his lab temporarily to Rice University in Houston, Texas, after Hurricane Katrina in 2005.

● **Don't dismiss your gut feelings.** Sudarshan Anand, a biologist at the Oregon Health & Science University in Portland, followed his

adviser from the Mayo Clinic in Rochester, Minnesota, to Johns Hopkins University in Baltimore. Although he acknowledges that some people rely on rational processes, such as cost-benefit analysis, to decide whether to relocate, that did not work for him. “Sometimes it's better to go with your instincts,” Anand says. He also recommends that early-career researchers avoid over-emphasising professional gains when mulling over a move. “You need to evaluate your personal situation,” he says. In his latest move, he had to consider his wife's employment prospects as well.

● **Line up a point person.** Brickman says that having a local coordinator on board to handle logistical details is invaluable. “We had an amazing administrator who was on the phone with my students and postdocs when their apartments fell through, or when we had any problems with animals,” he says. “That saved our bacon many times.” **P.S.**

represented more than 100 person-years of work — also proved cumbersome. First, he had to ensure that every line was duplicated in Edinburgh. Then he and his lab members arranged for the cell lines and reagents to be stored properly in liquid nitrogen and packed carefully into supercooled containers. The group loaded a truck with the cell lines and reagents and then flew to Copenhagen to meet it and ensure that the biological material was still stable.

Complicating matters further, the MRC Centre for Regenerative Medicine at the University of Edinburgh, where Brickman's lab was based, was moving to a new building at the time that the stem-cell centre in Copenhagen was under construction. When he learned that the opening in Copenhagen would be delayed, Brickman felt it best to move the bulk of his lab to temporary facilities in Denmark rather than to the new building in Scotland — even though he would later have to transfer again to the permanent lab.

Managing lab members in two sites also proved challenging. Brickman had landed a collaborative UK grant before he left, so he hired a new postdoc to work at his Edinburgh lab and continued to manage three lab members who remained there. He could not directly supervise his new recruits much of the time, and so missed out on day-to-day knowledge of how his Edinburgh lab functioned; he commuted between Scotland and Denmark weekly for three months but worked mainly in Denmark over the next two years. The protracted move, he says, may have delayed the publication of papers — an unfortunate result for his junior co-authors, although he says that the papers were eventually accepted into high-impact journals.

Despite all the snags, much went right, Brickman says. He credits administrative support in both Edinburgh and Copenhagen for the smooth relocation of his lab group. "All of my people were able to move both work and personal lives," he says. "None of them ended up homeless, despite moving to a new country where they didn't speak the language and in a city where it is almost impossible to find rental apartments." In the end, clearing the many logistical hurdles proved worthwhile, he says, because the new stem-cell centre's strengths outweigh the hassles that he underwent to join it.

There is no way around it — moving lab, whether within a university or to another country, is gruelling, stressful and likely to include disaster or catastrophe. Ultimately, however, no one can plan for everything, and adaptability is perhaps the most useful resource. "I am much more unflappable now," says Godbey. "The more extreme the situation, the more flexible you need to be." ■

**Paul Smaglik** is a freelance writer in Milwaukee, Wisconsin.

## TURNING POINT

# Daniel Carder

*Daniel Carder, director of the Center for Alternative Fuels, Engines and Emissions at West Virginia University (WVU) in Morgantown, was on a team whose work led to Volkswagen's admission that some of its diesel vehicles contained software able to sidestep emissions tests.*

### What does the centre do?

We have done vehicle-emissions testing and technology development for 25 years. We designed the first mobile diesel-fuel measurement systems, which use detectable carbon emissions to determine consumption. In addition to fundamental research, we produce open data on how new automotive technologies, such as clean diesel, prove in practice. We also try to make them more efficient.

### Can you describe your research?

I have bachelor's and master's degrees in mechanical engineering from WVU, and will complete my PhD this year. I am involved with the measurement and control of emission particulates related to diesel-fuel usage. My thesis work led to the adoption of US federal standards for particulate emissions in underground mine extraction. That technology controls highway and off-highway emissions.

### Did you expect to find any problems?

Quite the contrary. In 2013, we received US\$69,000 from the International Council on Clean Transportation to test the diesel emissions of two Volkswagen models. We expected to show that clean diesel fuel was doing a good job. We had seen successful demonstrations of the same type of technology in the bus and tractor markets and wanted to translate them for passenger-vehicle manufacturers.

### What were your first experimental results?

We believed that these systems would reduce emissions from 1,000 parts per million of particulates to 10–20 p.p.m. When we saw initial data for the Volkswagen vehicles, the first thing we did was scrutinize our work. Did we make a mistake with calibration? We double- and triple-checked our data and procedures. After several quality-control exercises, we were assured that our findings were valid. But it wasn't in our contract to find out why.

### Were the data made public?

Yes. Marc Besch, also a graduate student, presented the discrepancies in 2014 at a workshop in San Diego, California, attended by people from the US Environmental Protection



Agency, petroleum companies and engine manufacturers. Before we left the conference, we were contacted by Volkswagen asking about our techniques and data-collection methods. It seemed like a normal fact-finding mission.

### When did you realize that this was a big story?

I was in the lab on 18 September when the news broke. My hands were filthy from working on diesel engines. My phone was ringing continuously, but I didn't recognize the numbers — reporters were calling. We were blindsided.

### Can you describe the media attention?

Constant. I am the poster boy for why everyone should have media training. It's been trial by fire.

### How has the discovery affected your work?

You never know when routine research could have a major impact. And it has provided a good way to talk to our students about the quality and custody of data. It is satisfying and rewarding to be recognized for work behind the scenes.

### Do you have any concerns about the fallout?

We are the data collectors who will develop and refine new technologies. One concern is the perception that our objectivity could be compromised. We have strived to get industry to work with academia on emissions-control technology and policy issues because we believe that researchers cannot sit in ivory towers.

### What should the public know about your work?

As support for earmark funding has waned for centres such as ours with a mission that benefits the nation, we have shown that they have merit. It's difficult to keep ventures like this afloat without congressional support. ■

**INTERVIEW BY VIRGINIA GEWIN**

This interview has been edited for length and clarity.



# AGE PROGRESSION

*Family connections.*

BY SUSANA MARTINEZ-CONDE

**THEN:** Julia thought at first that the baby had slept through the night. She wanted to go back to sleep, but her breasts felt uncomfortable and milk-heavy, so she got up instead.

Inside the nursery, the familiar smells were soothing. The sweet buttery scent that was most intense over the soft spot on Rose's head; the vanilla fragrance from her bedtime lotion.

The room was too quiet.

She listened hard for the baby's breathing, willing her own lungs not to exhale, but only the ringing in her ears shattered the silence. She pawed at the light switch. Then she was standing over the crib, not knowing how she got there, tearing the blanket off her daughter's face.

**NOW:** Rose's wedding is next month, and Julia is helping with the flower arrangements. People say that she spends too much time obsessing about Rose, but she doesn't care to listen. Not a single one of her well-meaning advisers has walked in her shoes.

She regrets that Rose's father will not attend the simple but charming ceremony, but it can't be helped. He's been out of the picture since Rose was a little baby.

**THEN:** Julia watched the baby's head flop back and forth. She knew not to shake an infant but could not think of stopping; even less imagine the rest of her life after stopping. She kept shaking Rose until her husband, awoken by her screams, ripped the baby from her arms to attempt CPR. He kept at it, too, until the medics arrived and took Rose from him.

**NOW:** Julia never tires of looking at Rose's pictures. Her scrunched-up face minutes after she was born. The first toothless smile. The first shaky steps on chubby legs. Julia has documented every Halloween costume, every school graduation. The photos and home videos span more than two decades. She is amazed to think of how far she and Rose have come.

**THEN:** The grief counsellor at the hospital had arranged for the portrait artist, and Julia had been too numb to object. Now the girl lifted her eyes from her tablet and studied Rose. Julia tried seeing her baby, finally free from tubes and probes, through the



stranger's eyes, and failed. The little body on her lap was growing foreign to her, even as her shirt became damp with unsucked milk. Her eyes brimmed over and she thought it odd that her tears would still fall after she was no longer sobbing. She felt hollowed out and filled with liquid grief, spilling from her in every way.

**NOW:** The first portrait had been a miracle, a conduit that bridged the broken landscape of Julia's universe. Smiling at the memory, she opens the file in her computer and watches Rose's chest move up and down in her sleep, her eyelids fluttering as if about to wake.

**THEN:** Julia had resisted looking at the finished picture on the artist's tablet, but eventually relented, her mind seeking diversion from the horror on her lap. Then she couldn't stop watching. The baby's cheeks were flushed, and the slackness of her posture suggested sleep from which she might awake at any moment. "She looks alive," Julia said, acknowledging the young woman for the first time. Understanding the imploration in the mother's words, the artist took back the tablet and resumed her work.

Where Julia's husband had failed, and then the medics and doctors had failed, the artist succeeded. The newly animated portrait showed Rose's body moving in synch with her respiration, her rosebud mouth working as if dreaming about nursing.

**NOW:** When a baby dies, parents mourn not only the sweet infant that was, but also the spirited schoolgirl, the awkward teenager, the assured young woman that might have been, that *should* have been. They grieve

for their lost grandchildren. Julia no longer sorrows. She's not haunted by what ifs. The answers have unfolded in front of her for more than 20 years.

**THEN:** The hospital portrait, and the age-progressed reconstructions that followed, opened an irremediable schism in Julia's marriage. Her husband refused to partake in the immersive scenarios that later technological refinements allowed. His desire for another 'real life' child felt like a betrayal of Rose.

**NOW:** Julia's ex-husband has been happily married for most of Rose's life. He doesn't speak to his first wife or first-born child. Rose's half-siblings have never met her.

**THEN:** After her husband left, Julia consulted with the best virtual developers and age-progression biographers. She decided that Rose should have an ordinary life, not a fairy-tale one. Her daughter would experience the ups and downs, the minor disappointments and small joys that a regular child would. And she, her mother, would not run Rose's life as it evolved — she gave creative control to the company that she hired to age-progress Rose.

**NOW:** Julia is not a fool. She knows — how could she ever forget — that her baby died half a lifetime ago. *And yet.*

She remembers a college lecture about multiple worlds. The professor said that our universe splits every time we make a decision. When we park our car in the left spot rather than the right one, when we fail to set the alarm clock and sleep in instead of waking up early, when we stay in bed and don't check on the baby in the middle of the night, the universe breaks in half. There are as many universes as choices.

Who's to say that the virtual life Julia has created for Rose is not a window into the life that another Rose is living elsewhere, in a more benign reality? Julia likes to think it is.

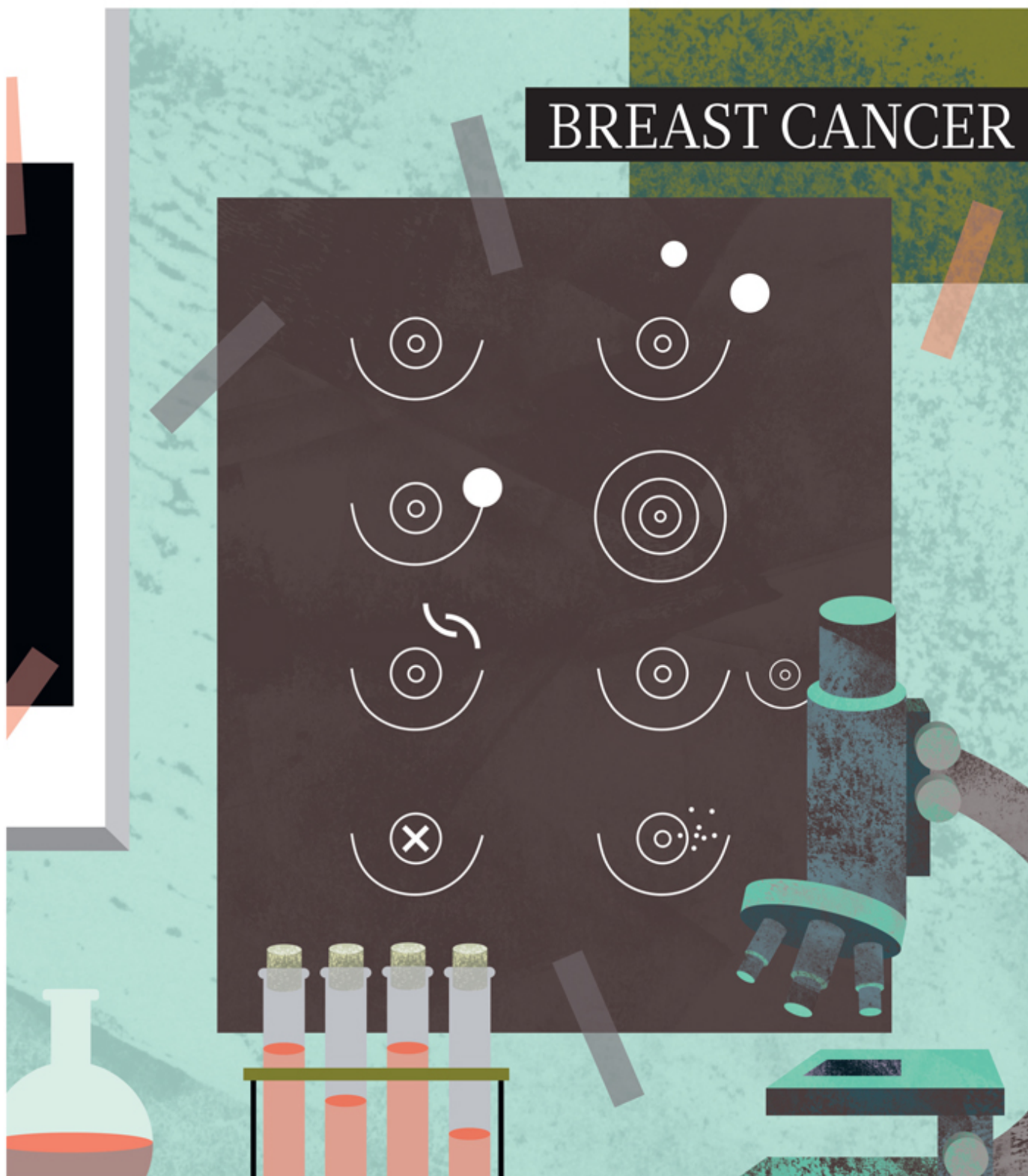
She goes back to the flower arrangements. Roses, of course. Her daughter is marrying young, but she may not wait too long to make Julia a grandmother.

She can't wait to meet her grandkids. ■

**Susana Martinez-Conde** is a scientist who writes for fun. In a parallel universe, she's a writer who dreams about doing science.

ILLUSTRATION BY JACEY

## BREAST CANCER



Produced with support from:

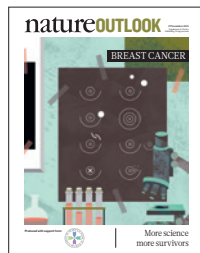


More science  
more survivors

# natureOUTLOOK

## BREAST CANCER

19 November 2015 / Vol 527 / Issue No 7578



Cover art: Elin Svensson

### Editorial

Herb Brody  
Michelle Grayson  
Chris Woolston  
Jenny Rooke

### Art & Design

Wesley Fernandes  
Andrea Duffy  
Denis Mallet

### Production

Karl Smart  
Ian Pope  
Mira Loufti

### Sponsorship

Yuki Fujiwara  
Yvette Smith

### Marketing

Hannah Phipps

### Project Manager

Anastasia Panoutsou

### Art Director

Kelly Buckheit Krause

### Publisher

Richard Hughes

### Chief Magazine Editor

Rosie Mestel

### Editor-in-Chief

Philip Campbell

**B**reast cancer is perhaps the most-studied malignancy in the world — and no wonder. Some 1.7 million women were diagnosed with the disease in 2012, making it a global priority. Researchers have made great strides in the treatment of some types of breast cancer (see page S102), but the battle continues on many fronts. Perhaps the most exciting area of research is immunotherapy (page S105), whereby scientists are attempting to harness the body's own immune system to fight and prevent malignancies. The success of biological drugs in the treatment of people with a specific tumour demonstrates the potential of targeted treatments (page S110). Meanwhile, researchers are using big data to identify new targets and treatment strategies (page S108). But not every cancer needs treatment, and the hunt is on for biomarkers that can sort the cases that demand action from those that are better left alone (page S114). Research on the interplay between environment and genes has illuminated the workings of the disease and helped to identify who is really at risk (page S116). Although many women try to protect themselves through regular mammograms, worries about false alarms and overdiagnosis have spurred efforts to reform screening to focus on the cancers that really matter (page S118).

However, each woman needs to decide for herself whether to be screened (page S104). And patients should have a say in the course of treatment. Some may want to go down the aggressive path no matter what the side effects, and others prefer to take the slow, cautious route. With many of the world's top minds working on their behalf, they should not feel alone in the fight.

We are pleased to acknowledge the financial support of the Medipolis Proton Therapy and Research Center, a part of the Medipolis Medical Research Institute, in producing this Outlook. As always, *Nature* retains sole responsibility for all editorial content.

**Chris Woolston**

*Contributing editor*

## CONTENTS

### S102 TIMELINE

#### **A tumour through time**

The long history of breast cancer

### S104 PERSPECTIVE

#### **The risks of overdiagnosis**

Why Alexandra Barratt may not get screened

### S105 IMMUNOTHERAPY

#### **Another shot at cancer**

Revisiting a type of treatment that was all but dismissed

### S108 GENETICS

#### **Big hopes for big data**

Genomic information could transform patient care

### S110 MEDICINE

#### **Eyes on the target**

The promise of antibody-based drugs

### S114 MOLECULAR BIOLOGY

#### **Marked progress**

Biomarkers could eliminate surgery

### S116 GENETICS

#### **Relative risk**

Cancer risk could be determined by the environment rather than genetics

### S118 SCREENING

#### **Don't look now**

Are mammograms worth the trouble?

### S120 BREAST CANCER

#### **4 big questions**

The areas scientists are still pondering

## COLLECTION

### S121 Breast cancer: Doubtful health

benefit of screening from 40 years of age

*Philippe Autier*

### S123 Precision medicine for metastatic breast cancer—limitations and solutions

*Monica Arnedos et al.*

### S135 BCL11A is a triple-negative breast cancer gene with critical functions in stem and progenitor cells.

*Walid T. Khaled et al.*

*Nature Outlooks* are sponsored supplements that aim to stimulate interest and debate around a subject of interest to the sponsor, while satisfying the editorial values of *Nature* and our readers' expectations. The boundaries of sponsor involvement are clearly delineated in the *Nature Outlook* Editorial guidelines available at [go.nature.com/e4dwzw](http://go.nature.com/e4dwzw)

#### CITING THE OUTLOOK

Cite as a supplement to *Nature*, for example, *Nature* Vol. XXX, No. XXXX Suppl., Sxx–Sxx (2015).

#### VISIT THE OUTLOOK ONLINE

The *Nature Outlook Breast Cancer* supplement can be found at <http://www.nature.com/nature/outlook/breast-cancer>. It features all newly commissioned content as well as a selection of relevant previously published material.

All featured articles will be freely available for 6 months.

#### SUBSCRIPTIONS AND CUSTOMER SERVICES

For UK/Europe: Nature Publishing Group, Subscriptions, Brunel Road, Basingstoke, Hants, RG21 6XS, UK. Tel: +44 (0) 1256 329242. Subscriptions and customer services for Americas – including Canada, Latin America and the Caribbean: Nature Publishing Group, 75 Varick St, 9th floor, New York, NY 10013-1917, USA. Tel: +1 866 363 7860 (US/Canada) or +1 212 726 9223 (outside US/Canada). Japan/China/Korea: Nature Publishing Group – Asia-Pacific, Chiyoda Building 5-6th Floor, 2-37 Ichigaya Tamachi, Shinjuku-ku, Tokyo, 162-0843, Japan. Tel: +81 3 3267 8751.

#### CUSTOMER SERVICES

Feedback@nature.com  
Copyright © 2015 Nature Publishing Group



# A tumour through time



## 200 MILLION YEARS AGO

Milk-producing skin glands evolve during the Late Triassic period in a group of egg-laying proto-dinosaurs called cynodonts. Instead of sweat or scent, these early mammarys produce a simple milk to supplement the diet of hatchlings. Over time, these glands grouped together under nipples and began responding to sex hormones such as oestrogen.

## ~2500 BC

A medical text from ancient Egypt contains the first known reference to breast cancer. It describes 48 surgical problems, including “bulging tumours on the breast”. The unknown author describes “swellings on [the] breast, large, spreading and hard: touching them is like touching a ball of bandages”. It would be some time before surgeons could offer anything beyond a diagnosis. The author’s take on treatment: “there is nothing”.

## ~1000

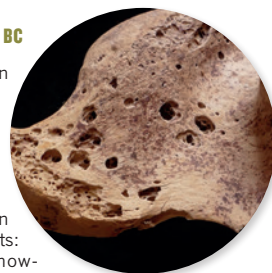
Doctors begin to embrace the possibilities — and limitations — of surgery for breast cancer. In eleventh-century Moorish Spain, the noted surgeon Abu al-Qasim al-Zahrawi writes that breast cancer could be cured when “complete removal [of the tumour] is possible, and especially when in the early stage and small. But when it is of long standing and large, you should leave it alone. I myself have never been able to cure such, nor have I seen anyone else succeed before me”.

## 1896

Emil Grubbe, an electronics enthusiast and medical student at the Hahnemann Medical College in Chicago, Illinois, assembles one of the earliest X-ray devices. A colleague remarks on the radiation burns on Grubbe’s hands and suggests that X-rays might be used against unhealthy tissue. The first recorded instance of radiation oncology occurs later that year when Grubbe’s machine is used to irradiate a breast carcinoma in a woman named Rose Lee. She reportedly received several hour-long treatments and died shortly thereafter.

## BOX OF BONES -2200 BC

During a 2015 excavation of the Egyptian necropolis Qubbet el-Hawa near Aswan, a team led by Egyptologists and anthropologists from the University of Jaén in Spain discovered a coffin with remarkable contents: bones of a woman, showing the tell-tale deformations of metastatic breast cancer (pictured). She died some 4,200 years ago, making her the first known victim of the disease.



## DEEP CUT 1590

French surgeon Barthélémy Cabrol suggests that advanced breast cancer could be cured by removing the underlying chest muscles along with the breast. Others will attempt variations on this idea for centuries, with often dismal results.



## KNIFE SKILLS -1894

In the United States, William Halsted and Willy Meyer independently develop radical mastectomy to treat advanced breast cancer. The operation removes the entire breast, the pectoralis major and minor muscles directly underneath, and the axillary lymph nodes from the armpit. The surgery offers people with advanced breast cancer the first serious hope of a cure.

## X-RAY VISION 1956

Robert Egan starts to develop effective mammography, an X-ray examination that can detect breast tumours that are too small to be felt. By the 1970s, mammography has become a popular screening test for women.

## DOUBLE STANDARD 1975

The National Surgical Adjuvant Breast and Bowel Project shows that surgery combined with chemotherapy works better against breast cancer than surgery alone. Combined treatments become the standard of care.

CLOCKWISE FROM TOP LEFT: 914 COLLECTION/ALAMY STOCK PHOTO; ARCHIVO PROYECTO QUBBET EL-HAWA; WELLCOME LIBRARY, LONDON; NATIONAL LIBRARY OF MEDICINE/SPL

*Breast cancer, one of the most common and deadly malignancies, has undoubtedly plagued humans since the dawn of our species. The history of the fight against the disease is one of lurching progress against a backdrop of misery. But recent decades have seen greatly improved treatments and increased survival. By Will Tauxe.*

### DRUG STOP 1977

The oestrogen-blocking drug tamoxifen (**pictured**) is approved in the United States as a treatment for advanced metastatic breast cancer. Today, tamoxifen is one of many hormone-blocking drugs used worldwide to treat — and in some cases prevent — certain types of breast cancer.

### BUSTED FLUSH 1995

The US Nurses' Health Study reveals that hormone replacement therapy (HRT) increases the risk of developing breast cancer. At the time, HRT was popular both to treat and prevent menopausal symptoms among post-menopausal women. Today, HRT is no longer routinely recommended for long-term use in post-menopausal women.

### LESS IS MORE 2002

Two large studies show that people with breast cancer live just as long after small lumpectomy surgery combined with radiation as they do after radical mastectomy. Further studies show that only a narrow 2-millimetre 'clean margin' of healthy tissue needs to be removed along with the cancer in a lumpectomy.

### TIMES CHANGE 2009

The US Preventive Services Task Force recommends that women should be offered a mammogram first at age 50, and then every other year after — a departure from previous advice to start annual screening at age 40. The change sparks debate about the balance between the harm of unnecessary treatment and the risk of undiagnosed cancers.

1990

Mary-Claire King and colleagues (J. M. Hall *et al. Science* **250**, 1684–1689; 1990) use samples of DNA from families with a history of breast cancer to establish a link between mutations in a tumour-suppressing gene she names *BRCA1* and an elevated risk of breast and ovarian cancer. The discovery changed thinking about genetic influences on cancer. Further research showed that mutations in another gene, *BRCA2*, could also increase cancer risk. Today, some women who test positive for these mutations — including actress Angelina Jolie (**pictured**) — choose to have their breasts removed to reduce their cancer risk.



2000

Charles Perou and colleagues (C. M. Perou *et al. Nature* **406**, 747–752; 2000) show that breast cancers can be grouped into clinical subtypes on the basis of mutations in their DNA. Analysis of tumour-cell DNA enables doctors to choose treatments that are more likely to be effective. The subtype of 'triple-receptor-negative' breast cancer is particularly difficult to treat because these cancers do not respond to signals from any of the breast growth hormones: oestrogen, progesterone and human epidermal growth factor 2.

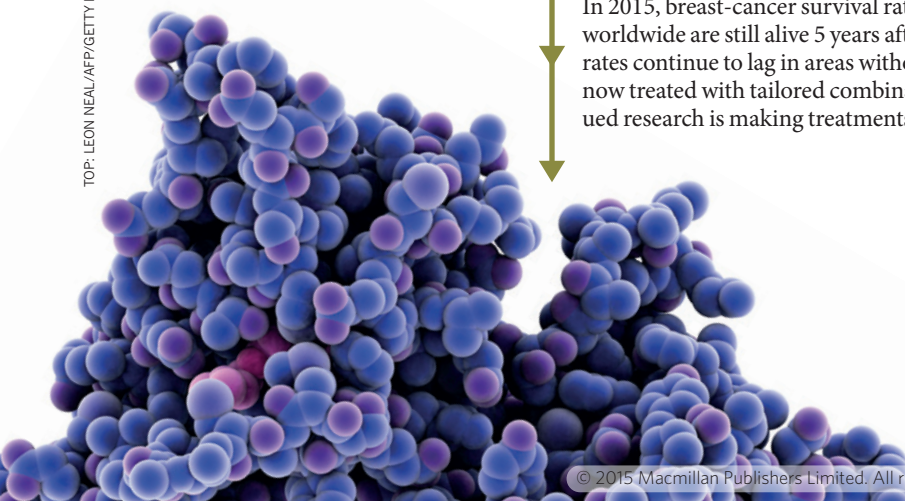
2013

In the case *Association for Molecular Pathology v. Myriad Genetics*, the US Supreme Court overturns molecular-diagnostics company Myriad's patents on the genetic codes of *BRCA1* and *BRCA2*. The court's decision states that "a naturally occurring DNA segment is a product of nature and not patent eligible", although tests to find specific harmful mutations are still considered patentable.

2015

In 2015, breast-cancer survival rates are at their highest ever. More than 6 million people worldwide are still alive 5 years after being diagnosed with breast cancer, although survival rates continue to lag in areas without reliable access to advanced medicine. Breast cancers are now treated with tailored combinations of surgery, chemotherapy and radiation, and continued research is making treatments more precise and minimizing side effects.

TOP: LEON NEAL/AFP/GETTY IMAGES; BOTTOM: LAGUNA DESIGN/SPL



## PERSPECTIVE



# The risks of overdiagnosis

Screening mammograms catch some cancers that pose little threat. Alexandra Barratt explains why she may decide to skip the scans.

Imagine a busy breast cancer clinic in a place such as Washington DC, London or my home city of Sydney, where women regularly have screening mammograms. In the waiting room sit women in their 40s, 50s and older who have breast cancer that was found by screening. They are scared and deeply uncertain about their future. But do they all need to be there? No, because some of them have been ‘overdiagnosed’<sup>1–3</sup> — they are about to have treatment for a breast cancer that would not have caused any health problems had it been left undetected and untreated.

In an ideal scenario, breast-cancer screening would find potentially lethal breast cancers before they have caused symptoms (such as a lump). Finding these cancers early would permit the best possible treatment and mean fewer breast cancer deaths<sup>2</sup>. This was the premise that underpinned screening programmes, but, increasingly, research shows that the picture is much more complicated.

Breast cancer takes many forms — some indolent and harmless, and some very aggressive and lethal that grow and spread rapidly. Because a screening mammogram is a snapshot in time, it is more likely to catch a slow-growing cancer than a fast-growing one<sup>3</sup>. In other words, it leads to overdiagnosis because of its tendency to detect the cancers that are unlikely to be harmful.

As a woman in my 50s, this is not just an academic issue. I have two options: I can have regular screening (every 2 years from age 50 to 74 is the recommended schedule in Australia), or I can choose not to. To decide, I need to consider the potential outcomes. One possibility is that my mammograms will always be normal and that will be that. Or perhaps one of my mammograms will not be normal and I will have further tests, which will show I do not have cancer. This is called a false positive. If I regularly go for screening for 20 years or so, there is more than a 40% chance that I will have one of these scares<sup>4,5</sup>. I would feel anxious, and it might take me some time to recover, but, on its own, I do not think the threat of a false alarm would stop me from being screened. The third possibility is that I am diagnosed with breast cancer. Like the women in the waiting room, I would be advised to have surgery (to remove the cancer or the entire breast), and probably radiotherapy, along with hormone therapy for 5 or more years.

So why would I even consider not getting screened? Am I not worried about dying of breast cancer? Yes, of course I am. But I’m also worried by the possibility that I could be seriously harmed by the treatment of a cancer that would never have affected my health.

Radiotherapy will increase my risk of heart disease, especially if the breast cancer is on the left, the same side as the heart. Cardiovascular disease killed my mother and three of my grandparents, so that is a big concern for me. I would almost certainly experience some of the common side effects of hormone therapy, such as tamoxifen, including mood problems, low libido and vaginal

dryness — and an increased risk of blood clots and stroke. Both my daughters would feel anxious and would, forever after, have a ‘family history of breast cancer’. Importantly to me, the psychological impact and physical risks of being treated for an overdiagnosed cancer would start straightaway. As I get older, I find that having good health and enjoying my life today becomes more and more important than the future.

At present, scientists and doctors cannot identify an ‘overdiagnosed’ cancer at the individual level because we do not have tests that can reliably distinguish between breast cancers that will progress to cause health problems and those that will not. Overdiagnosis can only be inferred from population-level statistics, so I could never be certain whether I had been overdiagnosed or not. I’m not sure that is a Pandora’s box that I’m willing to open. This is the conundrum at the heart of overdiagnosis, and the only way I can avoid it is by not getting screened in the first place.

I still might, however, decide to get screened because of the small chance that it could prevent me dying from breast cancer. By some estimates, screening 1,000 women regularly for 20 years might prevent 4 breast cancer deaths<sup>6</sup>. Some estimates are even lower<sup>4,7</sup>. That means that there is roughly 0.4% chance that screening will make the difference between me dying of breast cancer and not. Estimates vary, but I’m more likely to be overdiagnosed than to have my life saved — three to ten times more likely<sup>6,7</sup>.

To be clear, if I decide not to have screening, that does not mean I’m pretending breast cancer does not exist. If I notice a change in my breasts such as a lump, I will see my doctor without delay to have it tested. And if, in the future, the science about breast-cancer screening changes, I

will carefully consider the new information. For now, I will do what I can to minimize my risk of breast cancer by watching my weight, incorporating exercise into my life and moderating my alcohol intake.

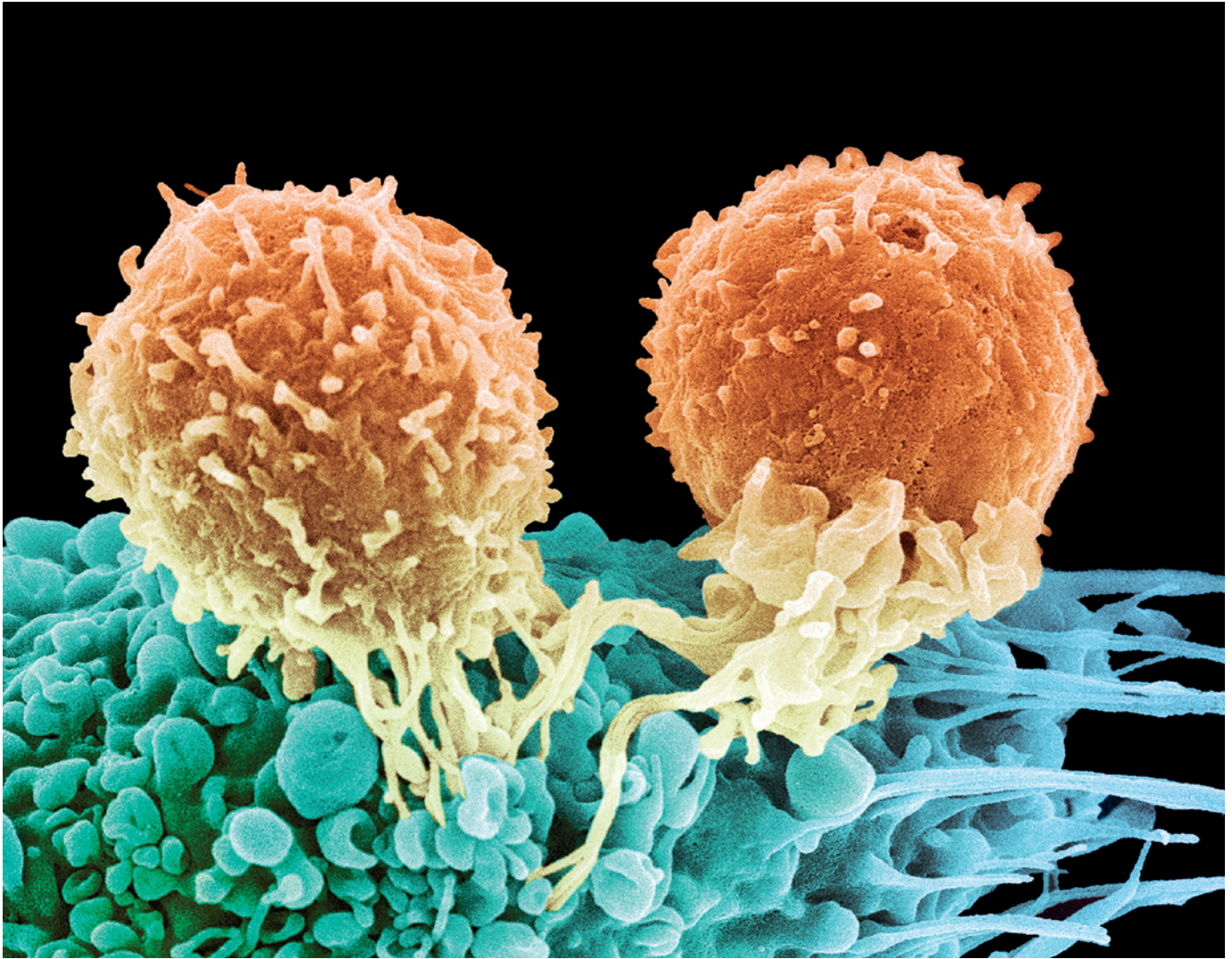
Whatever I decide, my decision should not determine that of anyone else. Women have different bodies, families, circumstances, preferences and fears. What matters is that each woman has access to the best information that science can provide to allow her to choose wisely what is best for her. ■

**Alexandra Barratt** is an epidemiologist at the University of Sydney, Australia.

e-mail [Alexandra.Barratt@sydney.edu.au](mailto:Alexandra.Barratt@sydney.edu.au)

1. Carter, J. L., Coletti, R. J. & Harris, R. P. *Br. Med. J.* **350**, g7773 (2015).
2. Esserman, L., Shieh, Y. & Thompson, I. J. *Am. Med. Assoc.* **302**, 1685–1692 (2009).
3. Welch, H. G. & Black, W. C. *J. Natl Cancer Inst.* **102**, 605–613 (2010).
4. Welch, H. G. & Passow, H. J. *JAMA Intern. Med.* **174**, 448–454 (2014).
5. Hersch, J. *et al. Lancet* **385**, 1642–1652 (2015).
6. Marmot, M. G. *et al. Lancet* **380**, 1778–1786 (2012).
7. Gøtzsche, P. C. & Jørgensen, K. J. *Cochrane Database Syst. Rev.* **2013**, CD001877 (2013).





White blood cells called T cells recognize and attack cancer cells as part of the immune response, which is boosted during immunotherapy.

## IMMUNOLOGY

# Another shot at cancer

*Targeting the immune system to fight breast cancer was all but dismissed in the 1990s, but the strategy is making a big comeback with the possibility of a breast-cancer vaccine.*

BY CHARLES SCHMIDT

While on her hospital rounds at the Walter Reed National Military Medical Center in Bethesda, Maryland, Elizabeth Mittendorf encountered a patient whose story is still fresh in her mind 14 years later. The woman had been successfully treated for breast cancer more than 15 years earlier, but the disease had returned. “And I wondered, how is it possible that someone beats breast cancer only to face it again?” recalls Mittendorf, now a surgical oncologist at the MD Anderson Cancer

Center in Houston, Texas. “To me this could only mean that this woman’s immune system had failed her.”

Mittendorf has since dedicated much of her career to breast-cancer immunotherapy — a field that is just starting to hit its stride. Immunotherapy drugs boost the body’s inflammatory response against malignant tumours. None have been approved for the treatment of breast cancer, and many uncertainties remain, but this is an undeniably exciting time. As of August, more than 40 clinical trials of breast-cancer immunotherapies are underway worldwide, and two of

them are in phase 3 — the final stage before regulatory approval can be sought. People with breast cancer already benefit from effective treatments, and 5-year survival rates for newly diagnosed cases top 90% in the United States. But drugs that enhance the immune system’s battle against malignancy might prevent recurrences altogether, says Mittendorf, the principal investigator in a phase 3 trial of a vaccine called NeuVax.

“What we hope is that immunotherapy will someday cure

➔ [NATURE.COM](http://NATURE.COM)

To read more on cancer immunotherapy, see: [go.nature.com/kvpgzl](http://go.nature.com/kvpgzl)



breast cancer,” says Mary Disis, an oncologist at the University of Washington School of Medicine, Seattle. “The immune system remembers cancer antigens, and it can seek out and kill off metastases anywhere in the body, including in the bones and the brain. We just have to figure out how to sustain that response before it’s exhausted.”

### A LONG HISTORY

The concept of cancer immunotherapy dates back more than a century. The bone surgeon William Coley, who worked at what later became the Memorial Sloan Kettering Cancer Center in New York, injected his patients with a killed bacteria vaccine during the late 1800s in the hope of stimulating the body’s defences. During the 1990s, physicians began treating people with cancer with high doses of interleukin-2 (IL-2) and interferon- $\gamma$  (IFN $\gamma$ ) — inflammatory cytokines released by infection-fighting white blood cells called T cells. Some people with cancer have lived for decades with the help of cytokine treatment, but because the inflammation that high-dose cytokines generate is systemic there can be life-threatening side effects, including vascular leakage and kidney damage.

A crucial breakthrough came in 1996, when James Allison, an immunologist at MD Anderson Cancer Center, and his colleagues showed that it was possible to amplify anti-cancer immunity by taking the brakes off a molecular checkpoint that would otherwise dampen the immune response<sup>1</sup>. The body relies on these checkpoints to regulate inflammation and limit the risk of autoimmune disease. But as they deliver this essential service, checkpoints interfere with the immune system’s efforts to destroy growing tumours. Allison’s research showed that blocking a checkpoint known as CTLA-4 located on T-cell surfaces enhances the immune response to cancer with fewer side effects than those brought on by IL-2 and IFN $\gamma$ .

In 2011, the US Food and Drug Administration (FDA) approved the CTLA-4 inhibitor ipilimumab for use in treating advanced melanoma. During phase 3 testing, people with the disease who were treated with ipilimumab lived an average of four months longer than those who went without the drug<sup>2</sup>. Some super-responders are still alive today.

While Allison targeted CTLA-4, other researchers were exploring the clinical possibilities of another immune checkpoint on T cells: programmed cell death protein-1 (PD-1). PD-1 binds to its ligand on cancer cells forming a complex called PD-1/PD-L1 and hiding the tumours from the immune system. Preventing the formation of these complexes has proved beneficial in cancer treatment. In December 2014, the PD-1 inhibitor nivolumab became the latest immune checkpoint therapy to gain FDA approval, specifically for the treatment of metastatic lung



**Trials of breast-cancer vaccines are underway, but the vaccines may work best when combined with other treatments.**

cancer<sup>3</sup>. European approval followed in April.

At first, there was widespread doubt that checkpoint inhibitors would be any use against breast cancers. The approach only works in tumours that have already been invaded by tumour-targeting white blood cells called tumour-infiltrating lymphocytes (TILs). Melanoma and lung tumours contain a lot of TILs, which makes them easy targets, but breast cancer tends to have relatively low levels of TILs. “So the thinking was that it wouldn’t respond as well to immunotherapy,” says oncologist Leisha Emens at the Johns Hopkins University School of Medicine in Baltimore, Maryland. “And since breast cancer was already being treated with effective drugs, it wasn’t associated with unmet medical needs in the same way that melanoma and lung cancer were.”

As checkpoint inhibitors make their way onto the market, attitudes have clearly shifted. “The entire medical community is awakening to an appreciation of their potential role in treating all cancers,” says Jill O’Donnell-Tormey, chief executive officer and scientific director of the Cancer Research Institute, a non-profit organization based in New York. “This is why you’re seeing all these clinical trials in breast cancer now.”

### TRIAL BY VACCINE

Of the more than 40 ongoing breast cancer immunotherapy clinical trials monitored by the Cancer Research Institute, roughly two-thirds involve vaccines. Breast-cancer vaccines

take a number of different forms: NeuVax, for instance, is derived from the cell-surface protein HER2, which some breast tumours have in large quantities, and is a target for the drug Herceptin (trastuzumab). Vaccines are also made from cancer-cell DNA, or entire cancer cells, and in some cases they are custom-made from a patient’s own white blood cells exposed to tumour antigens in the laboratory.

Whatever their origin, cancer vaccines are designed to stimulate a particular kind of anti-tumour immunity, specifically: type 1 immunity. Type 1 immune responses depend on CD4 T-helper cells that secrete highly inflammatory cytokines, such as IFN $\gamma$  and tumour necrosis factor- $\alpha$  (TNF- $\alpha$ ). In turn, these cytokines activate the CD8 T cells that go on to attack and kill cancer cells.

But vaccines may not work well enough as breast-cancer treatments by themselves. During the phase 1/2 trial of NeuVax, for instance, 89.7% of treated women achieved 5-year disease-free survival compared with 80.2% of women who did not receive the vaccine<sup>4</sup> — a result that some found discouraging. Mitten-dorf, however, argues that NeuVax was able to cut what would have been a 20% risk of 5-year recurrence in half, “which is a fairly impressive number that’s certainly of interest to patients.”

A phase 3 trial, called the PRESENT study, will randomize 700 women with early-stage breast cancer and low to intermediate HER2 expression to receive either NeuVax or an immune-stimulating chemical called granulocyte-macrophage colony-stimulating factor.

According to Mittendorf, if the trial reaches its endpoint of 3-year disease free survival, the FDA will consider the vaccine for approval — results are expected in 2018.

An increasing number of researchers, however, believe that the future of breast-cancer immunotherapy lies in giving vaccines and checkpoint inhibitors as combined treatments. In this way, vaccines will stimulate T-cell responses in the breast that checkpoint inhibitors can then amplify and sustain. Discussions about combining NeuVax with checkpoint inhibitors in future trials are ongoing. “I think this will be an efficacious strategy,” Mittendorf says.

## HITTING THE TARGETS

Meanwhile, the field is grappling with how to match people with breast cancer with the appropriate immunotherapy. The degree to which women with the most common form of breast cancer, oestrogen-receptor positive, will benefit from immunotherapy remains an open question, according to Carsten Denkert, a physician at the Institute of Pathology, Charité University Hospital in Berlin. Many oestrogen-receptor positive tumours, which make up 80% of all new diagnoses, are slow growing and respond well to existing hormonal treatments, such as tamoxifen or aromatase inhibitors.

Mittendorf points out that the number of women who die of oestrogen-receptor positive cancer that has stopped responding to existing therapies exceeds the number of women diagnosed with more aggressive types of breast cancer. The “challenge and opportunity,” she says, is to make oestrogen-receptor positive breast cancers better candidates for immune therapy, for instance, with combination treatments.

According to Christopher Heery, director of the Clinical Trials Group in the Laboratory of Tumor Immunology and Biology at the National Cancer Institute in Bethesda, Maryland, most immunotherapy trials look at highly aggressive triple-negative tumours, a much needed focus given that these cancers lack receptors for oestrogen, progesterone and HER2 — targets of existing cancer drugs. Debu Tripathy, who chairs the Department of Breast Medical Oncology at MD Anderson Cancer Centre, points out that the more mutated antigens that cancer cells carry, the more foreign they look to the immune system. And since triple-negative tumours express more mutated antigens than other breast cancer types, he says, they could be especially good candidates for immunotherapy.

Researchers hope to identify biomarkers that can help to predict which women with breast cancer will respond best to immunotherapy, but reliable candidates remain elusive. Checkpoint blockades in breast cancer have for the most part been limited to drugs that inhibit PD-L1. Examples include MPDL3280A, an

engineered monoclonal antibody manufactured by Basel-based Roche and currently in phase 1 testing for triple negative breast cancer, and pembrolizumab, manufactured by Merck in Kenilworth, New Jersey, which is in phase 2 trials with the same purpose in mind. It was initially thought that better responses would correlate with high PD-L1 expression levels; so much so that clinical trials have excluded women with breast cancer shown to be PD-L1 negative on screening. But PD-L1 expression is dynamic and varies not just between individuals, but also over time. Up- and downregulation of PD-L1 by cells is a normal response to excessive inflammation — cells upregulate it to reduce inflammation and downregulate it when the inflammation subsides.

Levels of PD-L1 may vary then depending on when samples are taken, and researchers still debate whether low PD-L1 levels should affect study enrolment. “Some people will say ‘You need high PD-L1 for checkpoint inhibitors to work and others will say ‘we ran a study and PD-L1 didn’t matter,’” Disis says. “The fact is that it’s just not a great biomarker.” According to Heather McArthur, a medical oncologist at Memorial Sloan Kettering Cancer Center in New York, there are other promising possibilities, including a marker for T-cell activation called ICOS, which predicts better responses to ipilimumab in patients with melanoma and a marker for T-cell proliferation called K167.

Another valuable biomarker, says Denkert, could be the amount of TILs in the tumour. oestrogen-receptor positive cancers tend to have low TIL levels, but that is not necessarily true of more aggressive malignancies, such as triple-negative and HER2-positive breast cancer. According to Denkert, about 25% of all

**“What we hope is that immunotherapy will someday cure breast cancer.”**

aggressive breast cancers are “lymphocyte predominant,” meaning that the number of TILs exceed the number of malignant cells. Another 25% of cancers have no TILs whatsoever, and the remaining 50% sit somewhere in between. Denkert’s research shows that a breast tumour’s TIL count is predictive of the response to chemotherapy — a high count predicts better responses — and he expects it to do the same for immunotherapy. “Tumours with zero lymphocytes will have only a small chance of responding to checkpoint blockade,” he says.

Boosting otherwise small amounts of TILs using vaccines could be a winning approach, Denkert says. But he thinks that in some instances, tumours characterized by low TIL levels might remain intrinsically invisible to the immune system even with this treatment because they do not express enough T-cell receptors. Denkert now plans to investigate that hypothesis in an upcoming clinical trial

sponsored by the German Breast Group, a network of the country’s academic research institutions.

Heery, however, cautions that not all TILs are equal. “Characterizing them is just as important as counting them,” he says. TILs could reflect type 1 immunity or type 2, which can suppress anti-tumour responses, Heery says. That is a crucial distinction, assuming that, as some research suggests, there tends to be a disproportionate number of type 2 TILs in breast cancer. Denkert agrees, but adds that approaches to discern type 1 and type 2 lymphocytes in tumour samples are in development. “Immunologists will say ‘the immune system is complicated and we have to look at the different types of immune cells,’ which is completely true,” Denkert says, “but we pathologists see this type of characterization as a second step that follows an initial effort to quantify the overall number of immune cells in the tumour. If we combine both worlds and accept different approaches, we can generate a more complete picture.”

As is the case with other experimental treatments, immunotherapies are being tested mainly in advanced, metastatic breast cancer. “Drug development is always done in the metastatic setting first to try to find the agents that work rapidly,” Heery says. The problem with cancer immunotherapies — especially vaccines — is that they can take up to three months to build up an adequate response, “so if sceptics don’t think they’re working well enough, it could be that we’re just testing them in the wrong setting.”

Ultimately, immunotherapies may have more success when used to treat early-stage breast cancer, and McArthur is one of the few investigators working in that setting. She selectively breaks tumour sections into fragments by freezing them with a tool that looks like a biopsy needle — the tiny tumour fragments are thought to attract a more robust immune response to cancers that are not highly immunogenic to begin with, she says. She is now testing this cryoablation approach in combination with ipilimumab in women with newly diagnosed oestrogen-positive cancer, independent of their HER2 status.

“This is an incredibly exciting time to be in oncology,” she says. “We’re seeing remarkable advances with immunotherapy in other solid tumours like melanoma and lung cancer, and I’m enthusiastic we’ll have success in breast cancer too. We’re on the cusp of a new era.” ■

**Charles Schmidt** is a freelance science writer in Portland, Maine.

1. Leach, D., Krummel, M. & Allison, J. *Science* **271**, 1734–1736 (1996).
2. Hodi, F. S. et al. *N. Engl. J. Med.* **363**, 711–723 (2010).
3. Kim, J. W. & Eder, P. *Oncology Suppl.* **3**, 15–28 (2014).
4. Mittendorf, E. A. et al. *Ann. Oncol.* **25**, 1735–1742 (2014).





ELIN SVENSSON

## GENETICS

# Big hopes for big data

*Technology is allowing researchers to generate vast amounts of information about tumours. The next step is to use this genomic data to transform patient care.*

BY JILL U. ADAMS

Adrian Lee has dedicated his career to studying breast cancer, which is to say he is actually tackling many different diseases at once. “No two breast cancers are the same,” says Lee, a pharmacologist and chemical biologist at the University of Pittsburgh in Pennsylvania. “Cancer is way more complex than we know.”

Lee is using genomic technology to fully describe cancers of the breast and apply that knowledge to guide treatment decisions for individual patients. “We can now analyse multiple variables from a single specimen, such as changes in DNA, changes in RNA and changes in methylation,” he says. “Genome-wide scans allow for better systems biology and allow us to learn what’s gone wrong in a particular tumour.”

Sequencing tumours is faster, cheaper and

easier than ever. With many researchers collecting sequence data and uploading these to public databases such as the The Cancer Genome Atlas (TCGA), opportunities to describe the many different cancers that arise in breast tissue are upon us. “The challenge used to be generating the data,” says Nicholas Navin, a geneticist at The University of Texas MD Anderson Cancer Center in Houston. “Those issues have been resolved. Now the challenge is data processing and data analysing — interpreting the mutations and communicating those to oncologists.”

At the University of Pittsburgh, researchers are working to link the molecular signatures of people with breast cancer to a host of clinical data, including demographic information associated with risk such as age, ethnicity and body weight. They are mining electronic health records for clinical correlates, treatment interactions and outcomes. “We’ve got a big haystack

and we’re trying to find the needle,” says Lee. “But we’re also trying to incriminate the needle, by linking it to lots of things.” Collecting all that data from patients’ electronic records adds up, Lee says. It takes infrastructure — Pittsburgh has already accumulated 5 petabytes, or 5 million gigabytes, which is enough data to overload around 40,000 new iPhone 6 devices.

Making the connection between the reams of data coming out of sequencing laboratories and the individual women fighting breast cancer takes big-time computing power. Big data needs researchers who are comfortable with statistical noise and those who are old hands at the iterative process required to create flexible computer programs.

## FROM DATA TO KNOWLEDGE

Big-data researchers take a large data set and look for patterns. The idea is to identify

mutations that can be targeted with drug treatment. It is the essence of personalized medicine: screen a patient's tumour for a set of biomarkers to choose the best treatment to fight the cancer. Big-data researchers believe that analysing the data of the thousands of tumours that have come before will reveal patterns that can improve screening and diagnosis, and inform treatment.

Lee and his colleagues have illustrated how big-data science led to a rethink of breast cancer<sup>1</sup>. They used two public databases — TCGA and METABRIC (Molecular Taxonomy of Breast Cancer International Consortium), which contain data on the entire set of genes, RNA transcripts and proteins of thousands of breast-cancer tumours — to parse out potential differences in the molecular signatures of breast tumours in younger compared with older women. Women who are diagnosed before the age of 40 tend to have worse disease: they are more likely to have later-stage cancers, poorer prognoses and worse survival outcomes than older women.

The team analysed tumour data from women under 45 years old, who were probably premenopausal, and women over 55 years old, who were probably postmenopausal. "We looked at everything you can look at," Lee says, including mutations in the genome, mutations in RNA, tumour gene expression, variations in the number of copies of certain genes and levels of DNA methylation. They found that tumours in premenopausal women follow a different playbook, especially in terms of gene expression.

As researchers find rarer and rarer mutations, the question of significance becomes more and more daunting, Lee says. He has just finished looking at a spreadsheet of 2,000 mutations. "One of them is the ER mutation," he says, referring to a mutation in the oestrogen receptor — a common mutation in breast cancers. "But how do I sift through the others? That's the fundamental problem."

One way to do it is to analyse the cellular pathways that the mutations affect. That means using algorithms developed to integrate all the collected molecular information and categorizing it into the common growth or cell-cycle pathways. Researchers can use this sorted information to describe tumours in terms of affected pathways rather than simply affected molecules. In one such effort, bioinformatician Josh Stuart of the University of California Santa Cruz developed a computational method that integrates a variety of genomic data sets with known cell-signalling pathways. "We know how gene circuits work in normal cells. Now we're asking, what got broken in this tumour cell?" Stuart says. "It's surprisingly successful."

Lee's group used the computational analysis PARADIGM in their study<sup>1</sup>. The approach proved particularly revealing for oestrogen-receptor-positive breast cancers in premenopausal women. The method demonstrated that although the individual molecules that showed

abnormalities varied, they often occurred within a particular set of pathways that signal for integrins — proteins involved in the formation of tumour-associated blood vessels.

The evident importance of integrins in the tumours of premenopausal women with oestrogen-receptor-positive breast cancers suggests that these molecules could be a therapeutic target. "There are integrin inhibitors out there," Lee says, and some of them have been tested in clinical trials.

### FROM KNOWLEDGE TO APPLICATION

As big-data researchers churn through large tumour databases looking for patterns of mutations, they are adding new categories of breast cancer. In 2012, two consortiums published papers on their data-driven approaches to breast-cancer genomics. The TCGA Network, made up of dozens of research institutions in the United States and Europe, came up with four overarching groupings of breast tumours based on genetic and epigenetic abnormalities<sup>2</sup>. They found that only three genes (*TP53*, *PIK3CA* and *GATA3*) were mutated in more than 10% of the samples, demonstrating that rare mutations are now an important part of breast-cancer typing. The METABRIC group, a consortium of UK and Canadian institutions, integrated genetic data — copy-number and gene-expression changes — with long-term clinical outcomes into 10 families of tumour types. Combined with clinical data, both these new groupings have the potential to allow oncologists to make better prognoses and treatment decisions<sup>3</sup>.

"We're still refining our approach," says Oscar Rueda, a biostatistician at Cancer Research UK's Cambridge Research Institute, which is part of the METABRIC effort. They are now fully sequencing the 2,000 samples used in the research. Rueda says that the hope is to identify driver mutations, which have a role in the initiation of cancer. "There are a hundred different mechanisms by which cells go bad," he says.

Big-data approaches may eventually reveal cellular pathways that had previously been overlooked. Avi Ma'ayan of the Icahn School of Medicine at Mount Sinai is working on a pathway database to create a resource for future potential targets. His effort comes under the umbrella of the National Institutes of Health Library of Integrated Network-based Cellular Signatures (LINCS), which uses data generated at institutions such as the Broad Institute of Massachusetts Institute of Technology. High-throughput labs at the Broad Institute test a host of drugs — both experimental ones as well as those with regulatory approval — on ten different cell lines to study how the drugs interact with cellular activity.

*"We know how gene circuits work in normal cells. Now we're asking, what got broken in a tumour cell."*

"You get a signature of what happens to cells," Ma'ayan says. "And signatures can be queried for new uses of drugs." If clinical researchers want to turn off a particular cellular pathway in cancer, they could use Ma'ayan's database to search for drugs that have that action.

### CLINICAL TRANSLATION

The next step is to apply the newly gained knowledge of actionable mutations to patient care. Research hospitals collect data on patients for their own care and to add to the knowledge base. At MD Anderson Cancer Center, for instance, people with a new cancer diagnosis are screened for a selection of cancer genes. "It's not the whole genome, but a panel of 200 genes with actionable mutations," Navin says. As research knowledge grows, so does the panel. In the past year, the original 200 genes have already expanded to 300, he says.

Navin's speciality is single-cell sequencing, which allows his lab to study tumour cells that are circulating in the blood. One might only collect 10 or 20 cells in a sample. "Previous analytic methods couldn't process such a small number of cells," he says. The single-cell approach opens the possibility that patients could be monitored over the course of treatment with a noninvasive test, such as a blood sample. Oncologists could then check if the tumour cells are responding to therapy or if resistance is emerging.

Big data intersects with the clinic in the form of I-SPY 2, a clinical trial of experimental breast-cancer drugs. "We're collecting real time data on patients," says Laura van't Veer, a molecular oncologist at the University of California San Francisco.

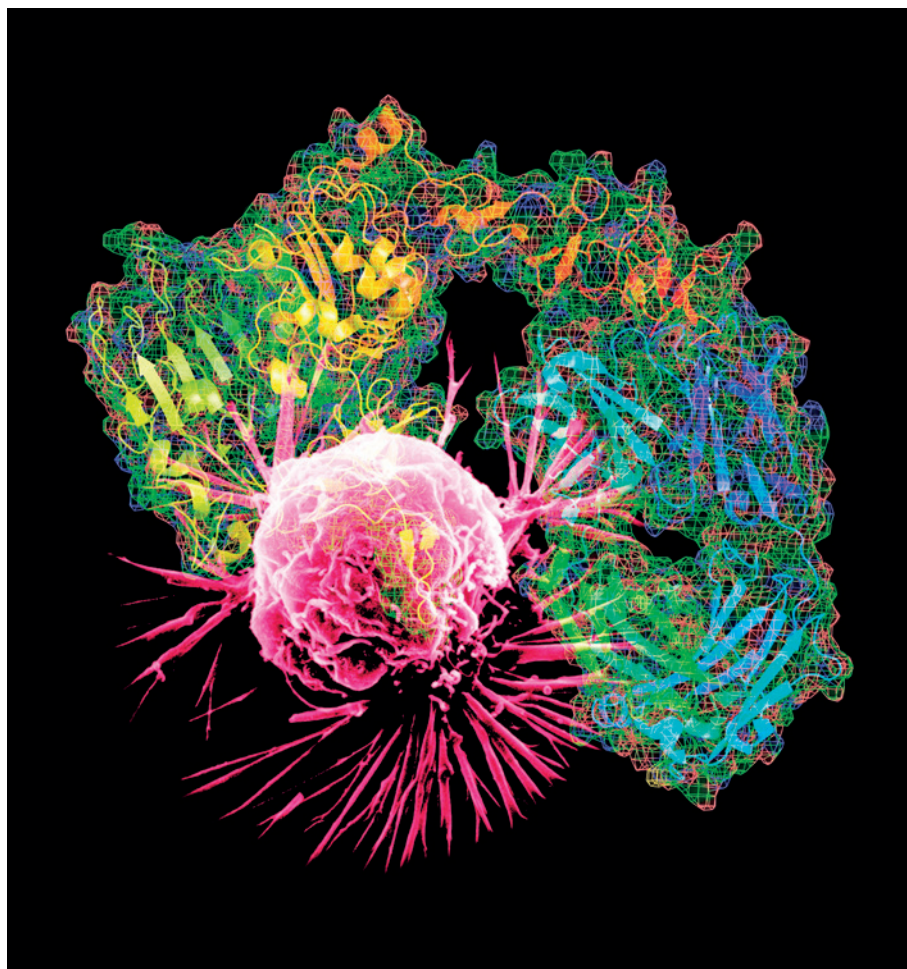
Patients are enrolled at diagnosis and, based on their tumour signature, placed into one of eight pre-defined types. The women are then treated with standard treatment and an experimental targeted drug, while van't Veer and her colleagues monitor which tumours respond to which targeted therapies. The goal is to evaluate biomarkers that improve response to targeted therapies. "With standard chemotherapy, we see 30–35% complete remission," says van't Veer. "Among our 8 subtypes, we sometimes get up to 50–60% remission."

Plenty of challenges lie ahead. A single tumour can host a baffling diversity of mutations, which change over time. Still, Ma'ayan remains the optimist. "The more money and effort we can throw at the problem, the more snapshots we can get. With better resolution, we can improve our understanding of the whole process," he says. "It's not infinite. Although it can feel like it." ■

*Jill U. Adams is a freelance science writer in Albany, New York.*

1. Liao, S. *et al. Breast Can. Res.* **17**, 104 (2015).
2. TCGA Network. *Nature* **490**, 61–70 (2012).
3. Curtis, C. *et al. Nature* **486**, 346–352 (2012).





Herceptin binds to a receptor called HER2 that is produced in excess by some breast cancer cells (pink).

# MEDICINE

# Eyes on the target

*A push to expand the success of a pair of antibody-based drugs is buying some women years of freedom from breast cancer.*

BY MICHAEL EISENSTEIN

Descriptions of a drug as revolutionary, transformative or a home run are usually reserved for press releases or presentations to investors. But oncologists are embracing such language to describe two drugs that allow them to offer some people with breast cancer a cure rather than a consolation.

The drugs are trastuzumab (Herceptin) and pertuzumab (Perjeta). Both are antibody-based agents that target the signalling protein HER2, which is produced in abundance in 20–25% of breast tumours. The high levels result in poorly controlled cell growth and proliferation. For decades, the

protein has been regarded as the hallmark of a dire prognosis, says oncologist Luca Gianni, at the San Raffaele Hospital Scientific Institute in Milan, Italy.

Today, many patients with HER2-positive tumours are essentially having their cancer eradicated by receiving a double-hit of targeted therapy before surgery. Even patients diagnosed with late-stage, metastatic disease — once seen as an imminent death sentence — are living much longer than ever before. In a few exceptional cases, the duration of these benefits can be remarkable. “We never use the ‘c-word’ with metastatic disease, but

I have one patient in my practice who has been in complete remission for 13 years,” says Shanu Modi, a medical oncologist at New York’s Memorial Sloan Kettering Cancer Center. “I think of her as a cured person who just comes by my clinic to visit every three months.”

These two agents exemplify the modern model of targeted therapy in oncology — give patients personalized treatments that selectively hit tumours based on their specific set of mutations, rather than conventional chemotherapy, which is broadly toxic to healthy as well as dangerous cells. Nevertheless, cancers will return in many people who receive prompt treatment. “I still don’t think we’re curing the vast majority of patients,” says Modi. New agents in the clinical pipeline could improve the effectiveness of these targeted agents and help doctors and patients to achieve more and longer-lasting victories, although some worry that soaring costs (see ‘Crippling costs’) will limit the reach of these next-generation therapeutics.

## A DYNAMIC DUO

Trastuzumab was the first HER2-targeting drug to reach the market, following a phase 3 clinical trial that showed that the drug improved the odds of survival by 20% in women with metastatic HER2-positive breast cancer<sup>1</sup>. Subsequent data showed that giving people the drug after the surgical removal of early-stage tumours cut the risk of the cancer returning in half<sup>2</sup>. “That’s an absolute home run,” says Elizabeth Mittendorf, a surgical oncologist at Houston’s MD Anderson Cancer Center. “We don’t see numbers like that often in oncology.”

In 2012, data from the CLEOPATRA clinical trial showed that oncologists could expect an even better return by pairing trastuzumab with pertuzumab<sup>3</sup>. Both drugs are antibodies that specifically bind HER2, but each recognizes a different site on the protein, and their combined effects (along with conventional chemotherapy) resulted in even greater tumour shrinkage and further improvements in prognosis for people with metastatic breast cancer. The combination prolonged median survival by well over a year relative to trastuzumab alone. “We give that combination to all of our metastatic patients as first-line treatment if we can,” says Modi, “and they’re being treated for three or four years on average.” In two further trials<sup>4,5</sup>, by using the trastuzumab-pertuzumab combination to shrink tumours before surgery — an approach known as neoadjuvant therapy — clinicians discovered that they could eradicate all traces of cancer from the breast and lymph nodes (known as a pathological complete response) in roughly half of patients.

That is not the same as a cure, however, and these trials are too recent to confirm

## ➤ NATURE.COM

To read more about breast-cancer research, visit.  
[go.nature.com/ztku8j](http://go.nature.com/ztku8j)



how durable the benefits are. Still, physicians are already seeing clear benefits in other domains. Mittendorf says she can now perform less-invasive operations that leave the breast largely intact, and neoadjuvant treatment can eliminate lymph node growths that previously required surgical removal. Oncologists are now eagerly awaiting the results from the recently concluded APHINITY trial, which measured survival and recurrence in patients given the combined therapy after surgery.

Other HER2-targeted therapies have reached the clinic, but without offering such a clear patient benefit. Lapatinib, for example, delays tumour progression by interfering with HER2 signalling, but also exhibits more toxicity and is generally less effective as a first line of treatment than its antibody-based counterparts. As a result, lapatinib is generally reserved for late-stage treatment of patients whose tumours acquire resistance to trastuzumab or pertuzumab. Importantly, this treatment may also be effective at limiting metastatic growth in the brain — a particular threat to patients who are HER2-positive (see 'Staving off a deadly invasion').

Another HER2-targeted drug called T-DM1 (Kadcyla) mitigates the severe toxicity associated with conventional chemotherapy agents by physically tethering the chemotherapeutic agent DM1 to trastuzumab. This restricts the toxic effect to HER2-expressing cells. "There's no hair loss and very little neuropathy," says Modi. "I think patients are enjoying a much better quality of life on T-DM1." As one of the first antibody-drug conjugates to reach the clinic, T-DM1 extends median survival by more than five months in patients with recurrent HER2-positive breast cancer relative to lapatinib.

A recently concluded trial called MARIANNE examined whether T-DM1 could replace trastuzumab and chemotherapy as the first line of treatment. Although less toxic than the standard drug combination, T-DM1 proved no more effective at delaying disease progression, and so will probably remain a second-line option. Gianni hopes that these results will not prevent clinicians from finding smart ways to incorporate this safe and generally effective drug into their regimens.

"T-DM1 was presented as the solution to all of our problems, and now it's being demonized as an expensive failure — but this is not so," he says. "It is simply a drug that requires some better thinking and a different approach."

***"There's no hair loss and very little neuropathy, I think patients are enjoying a much better quality of life on T-DM1."***

## TARGETED THERAPY

### *Crippling costs*

The life-extending advances in targeted therapy that are causing so much excitement for clinicians and patients with breast cancer are not cheap. In the United Kingdom, a course of T-DM1, for example, costs £90,000 (US\$138,000). The country's National Institute for Health and Care Excellence rejected routine coverage of the drug through the National Health Service (NHS) because of the high cost. For now, this drug is available through a special fund that provides access to medications that are not generally available through the NHS, although this fund is due to end next spring.

As costs soar, even the best drugs risk becoming bogged down in protracted negotiations. Consider the antibody-based agent pertuzumab. It was approved for presurgical, or 'neoadjuvant' use, by the European Medicines Agency in June, but cannot be used in this way until the various national health systems agree on a price. "Until recently, everything that had clear

benefits was approved and reimbursed," says oncologist Thomas Bachelot at the Centre Léon Bérard in Lyon, France. "But costs are really getting very high very rapidly, and there will be some tough decisions to make."

In the United States, private insurance generally covers these pricey treatments, but policies don't always cover the full costs, leaving many patients with crippling bills. In August, more than 100 clinicians signed a letter in *Mayo Clinic Proceedings*<sup>9</sup> to support a patient petition for vigorous pricing reforms, noting that every oncology drug approved by the Food and Drug Administration in 2014 cost upward of US\$120,000 per year — and that patients routinely shoulder up to a quarter of that cost. "We are negating therapeutic opportunities," says oncologist Luca Gianni at the San Raffaele Hospital Scientific Institute in Milan, Italy, "and this should not happen."

M.E.

### CUTTING OFF THE EXITS

The ability to target HER2 has been a life saver for many patients, but it is far from a complete victory. Many people who receive neoadjuvant treatment will not have their cancer eradicated. And although oncologists can keep cancer in people with metastatic disease at bay for years, this requires multiple rounds of therapeutic attack with a shifting arsenal of drugs. "The chance of having a durable response is much higher, and we're seeing patients with metastatic disease go through five, six or seven lines of treatment," says Nancy Lin, an oncologist at the Dana-Farber Cancer Institute in Boston, Massachusetts. "But the disease is generally not curable."

The mechanisms by which HER2-positive tumours acquire resistance to the drugs that once laid them low are poorly understood. Gianni sees the tumour environment, which contains a diverse mixture of cells with distinct mutational profiles, as part of the problem. "If 30% of the cells in a tumour overexpress HER2, that's a HER2-positive tumour," he says, "but many cells still do not express that target." Thus, killing off trastuzumab-vulnerable cells will still leave a large cancerous community. This highlights the value of generalized chemotherapy, and Francisco Esteva at New York University's Langone Medical Center suggests that this phenomenon may also be to blame for T-DM1's modest performance in MARIANNE. "If you target too much, the other

clones can escape," he says.

However, Thomas Bachelot, a medical oncologist at the Centre Léon Bérard, in Lyon, France, does not believe that this is the only mechanism by which tumours can recur after an initial therapeutic victory. "I do a lot of biopsies, and they always remain HER2-positive — they don't lose it," he says, adding that even if the tumour rebounds while patients are taking trastuzumab or pertuzumab, they still draw some benefit from those drugs. If HER2-targeted therapies are halted, he says, patients tend to fare even worse than if they had stayed the course. This outcome hints at other cellular pathways and processes that amplify or mitigate the effects of HER2-targeted treatment. These pathways might therefore serve as useful biomarkers to guide therapy. The trial data point to sub-populations with strong responses in both directions — one neoadjuvant trial<sup>4</sup> found that nearly 17% of patients had a pathological complete response from targeted drugs without any chemotherapy, a result that suggests that some patients could skip the most toxic components of treatment. Other tumours remain stubbornly unresponsive. "About 10% of our HER2-positive patients do not respond at all to trastuzumab and pertuzumab," says Bachelot. "There is this huge primary resistance, and we don't know why."

Another apparent pathway to resistance arises from hyperactivation of a signalling cascade known as the phosphoinositide

3-kinase (PI3K) pathway. Preliminary clinical studies suggested that everolimus, a drug that interferes with PI3K signalling, might increase the effectiveness of trastuzumab against metastatic breast cancer. Although results from two phase 3 trials proved disappointing, subsequent examination of the data revealed that everolimus may delay progression in patients with mutations that affect PI3K activity<sup>6</sup>. However, Modi believes that future studies should hit this pathway through alternate means. “You can get a few weeks improvement from everolimus, but with a lot of toxicity,” she says. “There have got to be better drugs for targeting PI3K.”

Her team is evaluating one such drug, BYL719, and has also seen promising results from another approach to bolstering HER2-targeted treatment. Tumour cells rely on a molecule called heat-shock protein 90 (HSP90) to manage the production of HER2, and Modi and colleagues have found that chemical inhibition of HSP90 can stall HER2-dependent tumour growth<sup>7</sup>. “In our first phase 2 trial, we combined an HSP90 inhibitor with Herceptin and saw a nearly 25% response rate from just these two agents, without any chemo,” she says. Her team is working with two different HSP90 inhibitors, including a compound that can be isotopically labelled so that the extent to which tumours are taking up the drug can be directly monitored.

Most breast tumours also show excessive activity by the receptors that respond to the reproductive hormones oestrogen and progesterone, and numerous studies suggest that these hormone receptors influence the response to HER2-targeted drugs. Data from dual-therapy neoadjuvant trials indicate that the pathological complete response rate jumps from around 50% to more than 80% in patients with hormone receptor-negative tumours<sup>5</sup>.

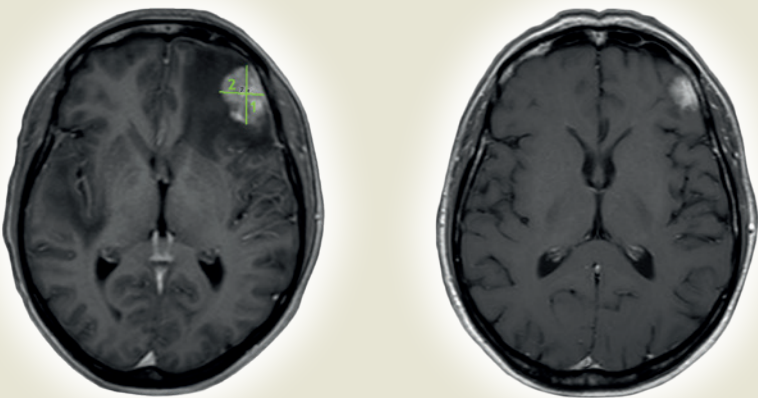
A growing body of data demonstrate that these two signalling pathways collaborate to promote growth<sup>8</sup>, suggesting that multiple hits may be necessary to limit the tumour's escape routes.

Ultimately, combinations of highly targeted treatments with more broadly active therapy regimens may hold the key to ensuring that no patient with cancer gets left behind, Gianni says. “We now have plenty of drugs that exploit different mechanisms of action,” he says. “The challenge is to use them in the optimal way.” For now, these targeted therapies are giving some patients the opportunity to live cancer-free, and giving many others the prospect of additional years of life — and a chance to think differently

**“We now have plenty of drugs that exploit different mechanisms of action, the challenge is to use them in the optimal way.”**

## METASTATIC TUMOURS

### Staving off a deadly invasion



**HER2-positive breast cancer can spread to the brain (green cross in left image); some patients have been shown to respond to targeted therapy to reduce the metastasis (right).**

Doctors and patients who wage a successful first-round bout with an HER2-positive breast tumour may subsequently find themselves grappling with a more fearsome opponent. “About half of those patients who have HER2-positive metastatic tumours develop cancer in the brain,” says oncologist Nancy Lin of the Dana-Farber Cancer Institute in Boston, Massachusetts. Part of the cause is biological, and Lin believes “there is probably some inherent predilection for these tumours to go to the brain”. However, the remarkable success of targeted treatments is itself a factor. The prolonged life of patients with metastatic breast cancer simply creates more opportunities for brain growths to form.

It remains unclear whether trastuzumab and pertuzumab are effective at hitting these metastases. Some researchers believe that the capillaries within the brain have too tight a seal for these relatively large drugs to penetrate. However, oncologist Thomas Bachelot of the Centre Léon Bérard in Lyon, France, thinks that this barrier is sufficiently leaky within tumours to grant these drugs access, and points to trial data suggesting that patients with stable brain metastases<sup>10</sup>

respond well to targeted treatment. Bachelot and colleagues have also found<sup>11</sup> that the small-molecule drug lapatinib, in combination with chemotherapy, controls brain metastases well. “You get very good results in terms of time to progression that are equivalent to what you see with disease outside the brain,” he says.

Radiotherapy approaches have also become more sophisticated; rather than subjecting patients to highly damaging whole-brain irradiation, clinicians can now selectively hit lesions while sparing healthy nervous tissue. People who once faced six-month life expectancies can survive five years or longer.

For years, the presence of disease in the brain automatically disqualified patients from being involved in trials. “It was often just part of the boilerplate language for eligibility,” says Lin, resulting in missed opportunities to identify regimens that might control or kill such growths. These restrictions are now being removed. Some companies are considering opening up trials to patients with untreated or progressive brain metastases, she says, a move that could quickly widen the battle against invasive breast cancer. **M.E.**

about their futures. “It’s completely changed the conversation in my office,” Mittendorf says. ■

**Michael Eisenstein** is a freelance science writer living in Philadelphia.

- Slamon, D. J. *et al.* *N. Engl. J. Med.* **344**, 783–792 (2001).
- Romond, E. H. *et al.* *N. Engl. J. Med.* **353**, 1673–1684 (2005).
- Baselga, J. *et al.* *N. Engl. J. Med.* **366**, 109–119 (2012).

- Gianni, L. *et al.* *Lancet Oncol.* **13**, 25–32 (2012).
- Schneeweiss, A. *Ann. Oncol.* **24**, 2278–2284 (2013).
- Slamon, D. J. *et al.* *J. Clin. Oncol.* **33**, abstr. 512 (2015).
- Modi, S. *et al.* *Clin. Cancer Res.* **17**, 5132–5139 (2011).
- Giuliano, M. *et al.* *Clin. Cancer Res.* **21**, 3995–4003 (2015).
- Tefferi, A. *et al.* *Mayo Clin. Proc.* **90**, 996–1000 (2015).
- Swain, S. *et al.* *Ann. Oncol.* **25**, 1116–1121 (2014).
- Bachelot, T. *et al.* *Lancet Oncol.* **14**, 64–71 (2013).





Some of the operations performed by breast-cancer surgeon Shelley Hwang may no longer be needed if better biomarkers for breast cancer are found.

#### MOLECULAR BIOLOGY

# Marked progress

*Reliable markers could eliminate surgery and radiation therapy for many women diagnosed with a type of cancer that often does not progress beyond its non-invasive form.*

BY HANNAH HOAG

**T**welve years ago, Mary Jane Lapinski had a routine breast-cancer screening mammogram at her local hospital in Baltimore, Maryland. The mammogram showed multiple specks in her left breast. Her physician called it ductal carcinoma *in situ* (DCIS) — an early-stage, non-invasive cancer of the milk ducts. A surgeon told her he could attempt a lumpectomy to remove the lesions, but he recommended a mastectomy — removal of the entire breast. “I kept thinking, this isn’t logical,” says Lapinski, who was 48 at the time. “It was mind-boggling that a non-invasive cancer carried the same or more aggressive treatment than an invasive cancer.”

The rogue cells in Lapinski’s breast occupied a diagnostic grey area. Some cases of DCIS advance to invasive breast cancer, metastasis

and death, but most do not. By current estimates, 20–30% of DCIS tumours will become aggressive within 20 years, says Shelley Hwang, a breast-cancer surgeon and researcher at Duke University School of Medicine in Durham, North Carolina. Still, most oncologists feel that it is best to remove the lesions and offer radiation treatment to stave off their progression.

The trouble is that oncologists cannot tell for certain which DCIS lesions will remain idle and which will turn deadly. Identifying breast-cancer biomarkers — molecules that can identify the pre-cancerous cells that are likely to progress to invasive cancer — could lead to better-informed decisions about treatment. Unfortunately, little is known about the natural history of DCIS. It is difficult to track the course of the disease because so many women undergo surgery. “If we can identify a subset of patients that are at risk of developing an

invasive cancer and only treat those, we would spare many women unnecessary treatment,” says Eileen Rakovitch, a radiation oncologist at Sunnybrook Health Sciences Center in Toronto, Canada.

#### SURGING DIAGNOSIS

Before the introduction of widespread screening mammography in the 1980s, DCIS lesions represented about 3% of breast cancers in the United States. They now account for nearly one-third of newly diagnosed breast cancers<sup>1</sup>. But detecting DCIS does not necessarily add much information about a woman’s future or overall health. “It is entirely possible to find cancers that don’t matter,” says H. Gilbert Welch, an internist and cancer epidemiologist at the Geisel School of Medicine at Dartmouth in Hanover, New Hampshire. Welch and Archie Bleyer, then at the University of

SHAWN ROCCO/DUKE MEDICINE



Texas Medical School in Houston, estimated that, in 2008, 70,000 US women received an early-stage breast-cancer diagnosis for lesions that would not have led to clinical symptoms, accounting for 31% of screening-detected breast cancers<sup>2</sup>.

Most women diagnosed with DCIS have a lumpectomy or mastectomy — or a double mastectomy — along with radiation therapy. But the benefits of such treatments are hard to find. A much-discussed observational study of more than 100,000 US women with DCIS found that women who had lumpectomies or mastectomies to treat DCIS had just a 3.3% chance of dying of breast cancer in the next 20 years, not much different than the risk to women in the general population<sup>3</sup> (2.7%, according to the American Cancer Society).

Ideally, women with DCIS would be tested to assess whether surgery is the best course of action. Although no such test is clinically available, physicians are starting to use biomarkers to predict the future of women who have already had surgery. One test — Oncotype DX DCIS Score, produced by Genomic Health in Redwood City, California — stratifies women who have had breast-conserving surgery for DCIS into low, medium and high risk for future cancer. The test evaluates the expression of seven cancer genes (including those associated with cell proliferation and hormone receptors) in tissue samples taken from breast biopsies.

Rakovitch validated DCIS Score in a retrospective study of women diagnosed with DCIS and who'd had breast conserving surgery. In work funded by a research grant from Genomic Health, she and her colleagues applied the test to tissue samples from 718 women<sup>4</sup>. "Women with an intermediate- or high-risk score had twice the risk of developing local recurrence compared to women with a low-risk score," says Rakovitch. She says that the assay can pick out some women who are at a high risk of recurrence, but whom doctors might have considered to be low risk based on patient history and tumour characteristics.

### SEARCHING FOR SIGNPOSTS

The search for a reliable measure to prevent surgery in the first place goes on. Any test, Hwang says, would probably involve a large array of markers that could be combined to form a cohesive picture. "We've taken the individual biomarkers as far as they can go and they're not giving us the answers we need," she says.

Thea Tlsty, a molecular pathologist at the University of California, San Francisco, and her colleagues have identified three proteins involved in cell proliferation that are associated with future aggressive breast cancers<sup>5</sup>. Tlsty's team found that of 1,162 women who



Ductal carcinoma *in situ* is a common type of non-invasive breast cancer.

had a lumpectomy for DCIS, those whose tissue was positive for all three biomarkers — COX-2, p16 and ki67 — had a 20% risk of developing an invasive cancer within 8 years. If they had none of the proteins, their risk dropped to 4%. "These markers indicate which pre-cancers are the baby, basal-like cancers, which are the most lethal and metastatic," she says. In unpublished research, Tlsty's group has subsequently identified several other potential biomarkers in proteins that coordinate cell death. Four prospective studies in Australia, the United States and the United Kingdom are further evaluating the trio of markers, Tlsty says.

Other biomarkers have also shown promise. Invasive breast cancers often stop the expression of tumour suppressor genes. One of those genes, called SYK, seems to be part of a genetic hub that determines which precancerous cells eventually metastasize. One study found that women who had altered expression of 55 genes that interact with SYK had reduced survival<sup>6</sup>.

The search for circulating markers for early detection has proved frustrating at times, says Jeffrey Marks, a cancer cell biologist at Duke University. Marks and his colleagues selected 90 blood-based biomarkers, but none were useful in distinguishing cases of breast cancer from benign controls<sup>7</sup>. "They're very difficult to validate in independent populations," he says.

Some researchers are looking for signals that might reveal which DCIS lesions are associated with an increased risk of developing future invasive breast cancer. Andy Beck, a computational biologist at Harvard Medical School in Boston, Massachusetts, and his team are examining patterns of genomic alteration. Using data from DNA profiles of invasive breast cancers catalogued with The Cancer Genome Atlas, the group identified genomic locations that are most frequently copied or deleted in invasive breast cancer lesions. "We're basically saying that if it's not present in invasive cancer

then it's not likely to be useful," says Beck.

In this case, the marker proved to be grimly robust. In a study of 271 patients, women with lesions that had extra copies in all three regions had a 17-fold higher risk of having a coincident invasive breast cancer compared with those women who had none<sup>8</sup>. The group is expanding the study to include about 20 chromosomal regions commonly altered in invasive breast cancer. In collaboration with Stanford University, Washington University and the Nurse's Health Study, the group is launching a study of 1,400 patients to predict the risk of recurrence or a subsequent invasive cancer over time.

Researchers have recognized that sheer genetic diversity within precancerous tissues may help to predict cancer formation and progression. As precancerous cells evolve and accumulate genetic and epigenetic alterations, they become more varied. Some studies have shown that diversity can predict progression. Marks is now studying the genetic diversity of the cells within DCIS lesions. In theory, if the DCIS has a more complex mosaic of cells, there is a stronger likelihood that one of them will develop into a more 'fit' cancer cell that can invade the surrounding tissue and metastasize.

Until scientists have a fuller understanding of which markers indicate an increased risk of developing invasive cancer, patients with DCIS will lack clarity about their future. For her part, Lapinski never went under the knife. Instead, she tracked down Hwang, who suggested that Lapinski join a three-month clinical study of the oestrogen-blocking drug tamoxifen. Lipinski was supposed to have surgery at the end of the trial, but she opted to forego the operation and continue with the tamoxifen, and, later, raloxifene. She checks in with Hwang twice a year for an examination and a mammogram. Although others see uncertainty in Lapinski's choice, she doesn't see it as a risky move. "Everybody has to make their own decision," she says. "It has to be comfortable for them." ■

**Hannah Hoag** is a freelance science writer in Toronto.

1. Esserman, L. & Alvarado, M. *Ann. Intern. Med.* **160**, 511–512 (2014).
2. Bleyer, A. & Welch, H. G. *N. Engl. J. Med.* **367**, 1998–2005 (2012).
3. Narod, S., Iqbal, J., Giannakeas, V., Sopik, V. & Sun, P. *JAMA Oncol.* **1**, 888–896 (2015).
4. Rakovitch, E. *et al. Breast Cancer Res. Treat.* **152**, 389–398 (2015).
5. Kerlikowske, K. *et al. J. Nat. Can. Inst.* **102**, 627–637 (2010).
6. Blacato, J. *et al. PLoS ONE* **9**, e87610 (2014).
7. Marks, J. R. *et al. Cancer Epidemiol. Biomarkers Prev.* **24**, 435–441 (2014).
8. Afghani, A. *et al. Breast Cancer Res.* **17**, 108 (2015).



Women born before 1940 have a much lower risk of developing breast cancer than their daughters.

## GENETICS

# Relative risk

*Mutations in BRCA genes predispose women to cancer, but outside influences shape the ultimate risk.*

BY MOISES VELASQUEZ-MANOFF

In 1990, geneticist Mary-Claire King forever transformed how we think about cancer with a single discovery: a mutation that dramatically increased carriers' risk of ovarian and breast cancer<sup>1</sup>. The gene *BRCA1* codes for a protein that is important in DNA repair. The mutated version impairs defences against tumours, increasing the lifetime risk of breast cancer in King's cohort to more than 80%, and the risk of ovarian cancer to as high as 40–65%. By comparison, the risk in the general population is 12% for breast cancer and 1.3% for ovarian cancer.

Four years later, another group identified a mutation in a second gene — *BRCA2* — that

also elevated the risk of these cancers, although by less. Mutations in the two *BRCA* genes are now thought to account for between 5 and 10% of all breast cancers, and 15% of ovarian cancers. These discoveries stand as landmark successes of the genomic era.

Although testing women for *BRCA* mutations is now commonplace for women with a family history of the disease (see 'Should all women be tested?'), the path between mutation and cancer is complex. Some studies show that the risk from *BRCA* mutations varies among different populations, suggesting that any particular woman's fate depends on more

than just her genes<sup>2</sup>. Among women who carry the mutation, additional factors — including exposure to oestrogen — may shape the risk of disease. Understanding the interplay between genes and the environment could illuminate the ultimate origins of breast cancer, possibly leading the way to new strategies for prevention and treatment.

## UNEQUAL RISKS

Some of the disparity in the risk from *BRCA* mutations is generational. One repeated finding is that, by age 50, mutation carriers born in the early twentieth century seem to have a lower risk of cancer than those born later<sup>3</sup>. The pattern suggests that outside influences interact with genes, and that something in the environment has changed in an unfavourable way. If researchers can figure out what those influences are, and why they have increased disease prevalence, maybe in the future they will gain new, less invasive tools to delay disease onset — and possibly prevent hereditary cancers altogether.

In 2003, King persuasively showed that the link between *BRCA* mutations and the risk of cancer varied with time. For Ashkenazi Jewish carriers born after 1940, the likelihood of developing breast cancer by age 50 was nearly triple that of women born before that date. "These people can be in the same family," says King, who is at the University of Washington, Seattle. "This is not genetic. The whole risk curve is getting shoved younger." This 'cohort effect' has been replicated by numerous researchers over the years, but its meaning is debated.

King attributes the generational shifts in *BRCA*-associated risks primarily to two trends: earlier starts to menses, and later first pregnancies. Women have been delaying first pregnancies more and more over the course of the past century. Meanwhile, girls now have their first menstruation about two years earlier than they did in the late nineteenth century.

Together, earlier menarche and later first pregnancy have increased the average woman's exposure to the sex hormone oestrogen, which is thought to promote tumour survival and growth. King believes this lengthened period of oestrogen exposure increases the risk of hereditary and non-hereditary cancers alike.

## WESTERNIZED HORMONES

Other researchers, however, think it is important to understand how our overall hormonal milieu may have changed over the past 100 years or so. Gillian Bentley, an anthropologist at Durham University in the UK who studies Bangladeshi immigrants, thinks that society-wide shifts could partly explain the increase of cancer during the past century both in *BRCA* mutation carriers and non-carriers.

One line of evidence is that the reproductive hormone levels of Bangladeshi immigrants vary according to when the women arrived in the United Kingdom. For those who came before puberty, adult hormone levels are similar

➤ **NATURE.COM**  
For more on cancer  
predisposition  
genes:  
[go.nature.com/xcxn1w](http://go.nature.com/xcxn1w)



to native-born Britons.<sup>7</sup> But if they arrived after puberty, their hormone levels remain suppressed relative to native Britons, but similar to levels of women in Bangladesh. Accordingly, South Asian immigrants who arrived as adults tend to develop breast cancer less often than native Britons. But their British-born children have a risk closer to native Britons. “They’re all from the same genetic background. We match them in terms of region of origin. And they move environments, and they look completely different,” she says. “What does that say about genes?”

Bentley suspects that childhood infections probably hamper the supply of reproductive hormone levels in people who grow up in Bangladesh. Because such infections were common throughout all societies in the past, it is possible that a similar scenario protected previous generations of western women from breast cancer. The lesson, however, is not to reinstate early-life infections, but to remember that genes interact significantly with the environment, Bentley says. “We need to understand the complexities,” she says, “and not be too simplistic in saying that genes determine your destiny.”

## IMPROVING THE ODDS

Joanne Kotsopoulos, a cancer researcher at the University of Toronto in Canada, is trying to help women who carry a *BRCA* mutation by identifying steps that they can take to protect themselves. Overweight women tend to produce extra growth factors and sex hormones, so staying slim may be one option. In a cohort of nearly 1,100 women, Kotsopoulos found that

**“There are lots of things you can’t change about your genetics.”**

*BRCA* carriers who had lost at least 4.5 kilograms between ages 18 and 30 had around half the risk of developing breast cancer by age 49 compared with carriers who did not lose weight.

By contrast, the use of oral contraceptives before age 20 correlates with a 45% increased risk of cancer in *BRCA1* carriers by age 40 (but not ovarian cancer)<sup>4</sup>. Because of the protective effect of oral contraceptives against ovarian cancer, she advises carriers to begin taking them once they reach 25.

What most excites Kotsopoulos, however, is exercise. She was first inspired by King’s 2003 study, which linked exercise in adolescence with a reduced risk of cancer later. In adults, exercise may lower breast-cancer risk by decreasing hormone and growth-factor levels. Regular physical activity in childhood can also delay menarche, shortening the period of oestrogen exposure. But Kotsopoulos suspects that exercise also helps in another way: by directly activating *BRCA* genes.

Many mutation carriers have one functioning *BRCA* gene. Exercise activates the functional copy sufficiently, Kotsopoulos thinks, to partly compensate for the non-functional copy<sup>5</sup>. In an,

## MUTATION SCREENING

### Should all women be tested?

Last year, Mary-Claire King, the geneticist who discovered the link between *BRCA* mutations and an increased risk of breast and ovarian cancer, and her colleagues argued in the *Journal of the American Medical Association* that all women should be screened for mutations in *BRCA1* and *BRCA2* (ref. 6). This approach, she says, would give women who might not otherwise know they are carriers a chance to consider proactive steps, including preventive surgery.

Some are sceptical of King’s proposal, however. Beverly Levine, a health policy expert at Wake Forest University School of Medicine in Winston Salem, North Carolina, says that the cost of such a programme per life saved would be forbiddingly high. She is concerned that universal testing might capture women with mutations who, because of little understood genetic or

environmental interactions, actually have a low risk of cancer. But oncologists will still recommend that they undergo prophylactic surgery. “Any time you have to involve a large number of people to prevent one case of disease,” Levine says, “there are unintended consequences.”

Such criticism does not faze King. Variants of unknown significance simply should not be included in the test results, she says. As for expense, the cost of these tests is falling rapidly. By her estimation, it might take 300 tests to save one life. And at a cost of US\$200 per test, she feels that the investment is worth the payoff. She is emphatic that younger women in particular need to get tested, even if they do not have breast or ovarian cancer in the family. “Women need to know,” she says. “Their genes aren’t going to go away.” **M.V.-M.**

as yet, unpublished study, she found that sedentary mutation carriers have less *BRCA* gene expression than more active carriers.

These are just associations, she stresses. They require confirmation in prospective research, and further testing in intervention studies — they are not meant to replace preventive surgery for mutation carriers. Rather, Kotsopoulos’ goal is to provide evidence-based advice to the significant number of patients who decline surgery. Eventually, however, understanding these ‘soft-risk’ modifiers may lead to new, less-invasive treatments, including drugs that mimic the *BRCA*-activating effects of exercise — an approach Kotsopoulos has begun testing in people.

## A CARCINOGENIC MICROBIOME?

Mysteriously, however, even when researchers control for risk modifiers, such as body fat and physical activity, they still cannot completely abolish the cohort effect. All else being equal, older women still seem to have a lower lifetime risk of breast cancer than younger women. To explain the increase over time of breast cancer generally, some researchers are turning their attention to the human microbiome.

Susan Erdman, a microbiologist at Massachusetts Institute of Technology in Cambridge, suspects that diet- and antibiotic-driven shifts in our microbial communities have, by encouraging chronic inflammation, increased the risk of breast, ovarian and prostate cancer in westernized populations.

Erdman has proven the basic concept in animal models. In mice, a junk food diet can increase the risk of these malignancies, apparently by altering the microbiome and

lowering the immune system’s ability to halt inflammation.

In the future, Erdman says, therapeutically targeting the microbiome could help to lower cancer risk. “There are lots of things you can’t change about your genetics,” she says. “But there’s lots you can change about your interaction with microbes.”

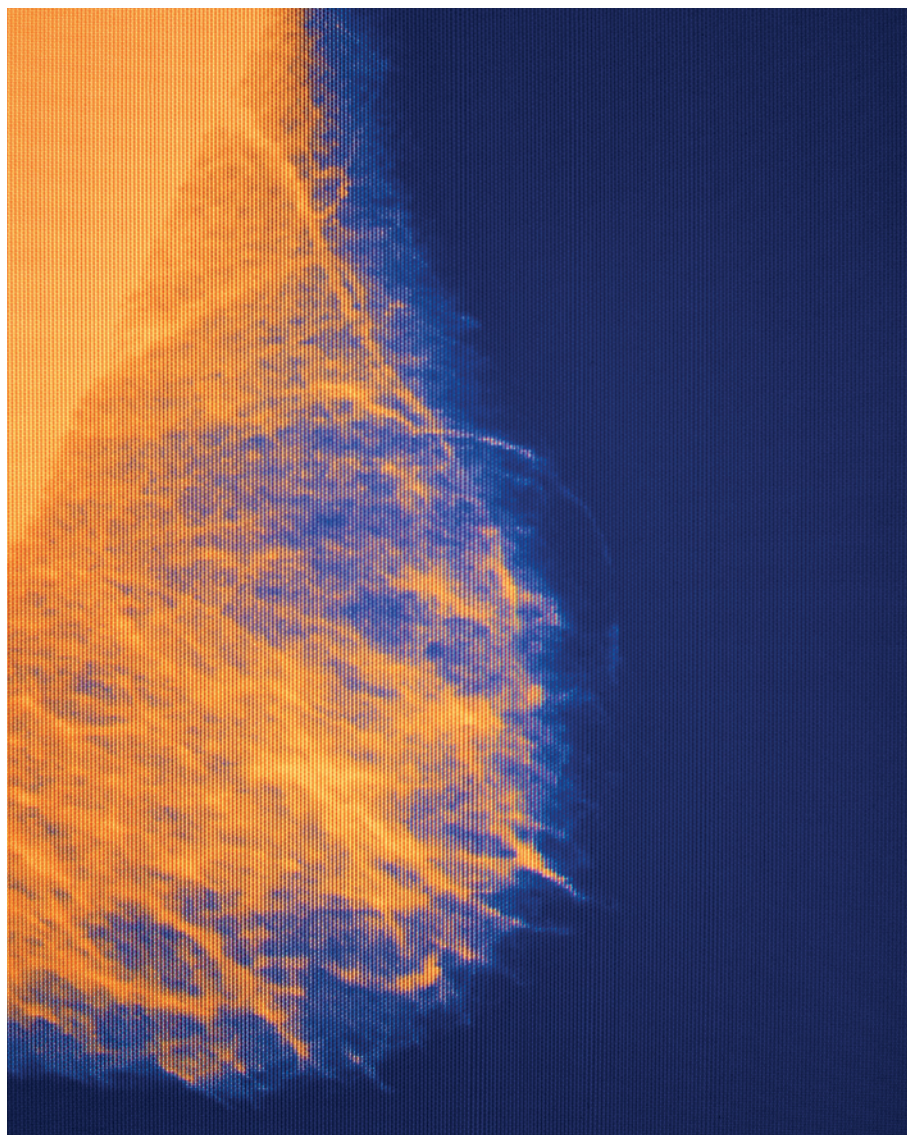
Looking forward, scientists are trying to understand how cancers linked with *BRCA1* and *BRCA2* mutations interact with non-hereditary factors (as well as with other genes). Perhaps a better understanding of this interplay will allow oncologists to provide more precise, individualized risk assessments that reduce the number of unnecessary surgeries. And maybe one day carriers will know how to reliably, and meaningfully, alter their risk of developing breast cancer with interventions other than surgery.

For now, King says that the options for women carrying a *BRCA* mutation remain limited. They can either live with increased risk and fear, or they can get a preventive operation. “I wish there were interventions that were safe that we could use,” she says. “We don’t presently have a non-surgical intervention.” ■

**Moises Velasquez-Manoff** is a freelance science writer in Berkeley, California.

1. Hall, J. et al. *Science* **250**, 1684–1689 (1990).
2. Rebbeck, T. & Domcheck, S. *Breast Can. Res.* **10**, 108 (2008).
3. King, M. et al. *Science* **302**, 643–646 (2003).
4. Kotsopoulos, J. et al. *Breast Can. Res. Treat.* **143**, 579–586 (2014).
5. Pettapiece-Phillips, R., Narod, S. & Kotsopoulos, J. *Can. Causes Control* **26**, 333–344 (2015).
6. King, M., Levy-Lahad, E. & Lahad, A. *J. Am. Med. Assoc.* **312**, 1091–1092 (2014).





Mammography is used to examine the soft tissues of the breast in order to screen women for cancer.

## SCREENING

# Don't look now

*Mammogram screenings are an established part of women's health care, but are they more trouble than they are worth?*

BY EMILY SOHN

Back in 1980, Canadian researchers decided to tackle a pressing question: how many women could be saved with breast-cancer screening? At the time, many researchers suspected that the relatively new technology could prevent breast-cancer deaths by detecting tumours before they had a chance to grow. The truth proved to be complicated — so complicated that researchers are still trying to understand it 35 years later.

The Canadian trial enrolled nearly 90,000 women from across Canada and randomly assigned them to receive either mammograms and physical breast examination or just breast examinations<sup>1</sup>. By 1990, there were enough data to begin analysis. But the researchers were confused by what they were seeing: the women who had received mammograms were not living any longer than those who had not. The team was so perplexed that it decided to delay publication, says Cornelia Baines, a physician epidemiologist at the University of Toronto in Ontario. “We

thought, ‘if we wait another two years, maybe a benefit will begin to emerge,’” she says. “After 25 years, it still hasn’t emerged.”

Health campaigns often give women a simple message: mammograms save lives. But the data behind that message is so murky that experts continue to disagree about who really needs to be screened. Various organizations offer clashing recommendations that range from the American College of Obstetricians and Gynecologists suggesting that regular mammograms should start at age 40 to the Swiss Medical Board proposing the elimination of routine screening altogether. At the same time, scientists continue to sort through the data from multiple trials, enlist the power of computer models, conduct epidemiological analyses, and take a closer look at the physical and emotional tolls of cancer treatments to address a vexing problem: are mammograms helping women, or hurting them?

## INVADER DETECTION

Decades after it started, the Canadian trial illustrates the essence of the mammography conflict — overall, women who were tested did not live longer, even though screening did uncover many cases of invasive cancer. Those diagnoses set off rounds of chemotherapy, radiation and surgery. But some of the women were already too sick to save, and others had cancers that could have been ignored: the authors estimated that 22% of the invasive cancers detected by mammograms were unlikely to affect a woman's health, a phenomenon known as overdiagnosis.

The Canadian trial is not the final word, however, and other studies have suggested that widespread screening with mammograms may reduce the overall death rate from breast cancer, although by how much is another question that scientists continue to debate. One review of the literature concluded that mammograms can reduce deaths by 20% (ref. 2), whereas another that looked at 7 randomized trials put the number at 10–15% (ref. 3).

After systematically reviewing studies that compared similar groups of women who received different rates of screening, a team led by Russell Harris, an epidemiologist and cancer-prevention researcher at the University of North Carolina School of Medicine in Chapel Hill, and his colleagues estimated that regular screening for women between the ages of 50 and 69 could reduce the death rate by about 10% (ref. 4). The benefits are smaller for women in their 40s, Harris says, because the cancer is less common.

For many women, the chance to reduce their risk of dying from breast cancer, even by a small amount, with a fairly simple procedure makes mammography seem worthwhile. But there are reasons to be cautious, says H. Gilbert Welch, an internist at the Geisel School of Medicine at Dartmouth in Hanover, New Hampshire. What a 20% benefit actually adds up to is a reduction from 5 deaths out of 1,000 women over a 10-year period to 4 deaths.

Proving that mammograms, along with other screening tests, extend lives has been notoriously difficult, Welch adds. One reason is that randomized trials — the gold-standard for assessing any kind of medical procedure — need to follow tens of thousands of people for many years, often requiring “heroic efforts” to keep tabs on mostly healthy people, he says. Among other challenges, it is hard to standardize the procedure. Results vary depending on who is reading the images, how hard they look for tumours and what kind of abnormalities get flagged for a call-back.

As trials have marched on over the decades, the landscape of breast cancer has shifted. A rising awareness makes women much more likely to detect lumps and get them checked out without the need for mammograms. Meanwhile, new, targeted treatments allow oncologists to extend the lives of women with advanced cancers who would have died owing to lack of options when trials began. “As we become better able to treat advanced stages of breast cancer,” Welch says, “it becomes less important to find early forms of it.”

### TOO MUCH KNOWLEDGE

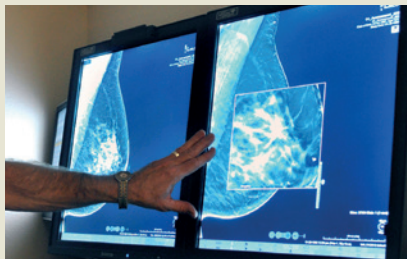
Determining the value of mammograms must take into account the problem of overdiagnosis. Peter Gøtzsche, director of the Nordic Cochrane Centre in Copenhagen, and author of *Mammography Screening: Truth, Lies and Controversy* (CDC, 2012), led a study that suggested that one-third of Danish women are overdiagnosed<sup>5</sup>. In other countries, including the United Kingdom, Australia and Sweden, the overdiagnosis rate as a result of introducing screening programmes is suggested to be in excess of 50% (ref. 3).

But not everyone agrees that too many women are being screened. Etta Pisano, a radiologist at the Medical University of South Carolina in Charleston, argues that the real problem is over-treatment that occurs after the mammogram. She compares mammograms to metal detectors at airports. Security personnel want to take a close look at anything that looks remotely suspicious because they “want to find all the guns,” she says. But when the ‘guns’ are tumours, which can behave unpredictably, the path forward remains a judgement call.

One major problem, Welch suspects, is that technological improvements (see ‘Fewer places to hide’) have given screening tests the power to detect more small and slow-growing cancers — twice as many in 2008 as in 1976, according to one study<sup>6</sup>. But the tests are still not finding more of the early stages of aggressive, fast-growing and lethal tumours that often appear between mammograms. If routine screening tests were catching the worst cancers early, Welch says, the data should show progressively fewer late-stage diagnoses as the number of early-stage diagnoses goes up, along with a reduction in death rates. But when he and his colleagues looked at data on 16-million

## NEW TECHNOLOGY

### Fewer places to hide



Mammograms are not the only way to screen for breast cancer. Ultrasound and magnetic resonance imaging have been available for years, particularly for women at high-risk of the disease. And a wide range of new technologies is making ever-smaller tumours increasingly easy to spot.

Tomosynthesis (pictured) takes X-ray slices of the breast to create a 3D image. According to a 2014 analysis<sup>9</sup> of more than 450,000 breast examinations from 13 centres around the United States, combining tomosynthesis with digital mammography (which uses computers instead of X-ray film) found more cancers with fewer false

positives than did mammography alone. But tomosynthesis delivers nearly twice as much radiation, says Etta Pisano, a radiologist at the University of South Carolina in Charleston. And comparative data are lacking. Pisano is planning a trial that would randomly assign 70,000 women to have either tomosynthesis or mammography to obtain the first accurate assessment of which technique helps women the most.

Further down the line, technologies in development include breast computed tomography scanners that are similar to tomosynthesis, but capture a more complete 3D image without as much radiation; phase-contrast imaging, which would show more detail of the interface between tumours and the surrounding tissue; and a bra that uses thermodynamic sensors to look for temperature fluctuations that might indicate signs of cancer. For now, the US Preventive Services Task Force has determined that more research is needed before it can recommend any of the emerging technologies. **E. S.**

American women living in nearly 550 counties that varied in mammography rates from less than 40% to nearly 80% they found that neither is happening<sup>7</sup>. Mammograms may do a good job at finding cancers, he says, but they do not seem to save many lives.

Mammography also carries risks. Besides the low-dose radiation from the X-rays, women with abnormal results often go on to be scanned with even more powerful imaging machines, and sometimes this is followed by biopsies, chemotherapy, surgery and other painful or risky procedures that can cause infections or worse. Radiation therapy as a treatment for breast cancer increases deaths from heart disease by more than 25% and from lung cancer by nearly 80% — a big risk for a woman who may not need to take it<sup>8</sup>. The consequences are not just physical. Harris notes that being labelled as a ‘cancer patient’ can turn lives upside down. And although the emotional fallout has been poorly studied, interviews with women suggest that worrisome mammogram results are common sources of anxiety, intrusive thoughts, insomnia and other kinds of distress that can endure for a lifetime, he says.

There are also financial concerns. In one study, researchers totalled the money spent on breast cancer in the United States between 2011 and 2013. Using data collected by a health insurer on more than 700,000 women<sup>8</sup>, the authors found that US\$4 billion was spent each year on care for breast cancer as a result of false positives and overdiagnosis, including

extra X-rays and biopsies. “It is a lot of money and it gives you a sense that this is not a small problem,” says co-author Kenneth Mandl, a biomedical informatics specialist at Harvard Medical School in Boston, Massachusetts.

Meanwhile, a strong narrative of breast cancer survivorship endures, partly because it is impossible to know after treatment ends whether it was needed in the first place. “That is the screening paradox,” says Gøtzsche. “The more healthy women you give an overdiagnosis to, the more they will tend to be happy because they think, ‘Screening saved my life.’”

It is an enduring irony that plagues the daily decisions women make about their health. “I believe mammography helps a few women,” says Welch. “But it’s a very few and it comes with tremendous human costs.” ■

*Emily Sohn is a freelance journalist living in Minneapolis, Minnesota.*

1. Miller, A. B. *et al.* *Br. Med. J.* **348**, g366 (2014).
2. Independent UK Panel on Breast Cancer Screening. *Lancet* **380**, 1778–1786 (2012).
3. Gøtzsche, P. C. & Nielsen, M. *Cochrane Database Syst. Rev.* **19**, CD001877 (2011).
4. Harris, R., Yeatts, J. & Kinsinger, L. *Prev. Med.* **53**, 108–114 (2011).
5. Jørgensen, K. J., Zahl, P.-H. & Gøtzsche, P. C. *BMC Women's Health* **9**, 36 (2009).
6. Bleyer, A. & Welch, G. N. *Engl. J. Med.* **367**, 1998–2005 (2012).
7. Harding, C. *et al.* *JAMA Intern. Med.* **175**, 1483–1489 (2015).
8. Ong, M.-S. & Mandl, K. D. *Health Aff.* **34**, 576–583 (2015).
9. Friedewald, S. M. *et al.* *J. Am. Med. Assoc.* **311**, 2499–2507 (2014).



## BREAST CANCER

## 4 BIG QUESTIONS

*Although treatments for breast cancer have come a long way over the past few generations, researchers are still puzzling over some tough questions.*

BY CHRIS WOOLSTON

## QUESTION

## WHY IT MATTERS

## WHAT WE KNOW

## NEXT STEPS

1

**Which cancers need to be treated?**

Treatment for breast cancer can be disfiguring, expensive and painful. Identifying the tumours that actually pose a threat can focus therapy where it is needed, while saving millions of women from potentially harmful interventions.

Roughly 15–30% of tumours detected by screening are unlikely to cause trouble if left alone. Ductal carcinoma *in situ*, — a common cancer of the milk ducts — rarely turns invasive, but women with the disease often undergo surgery (see page S114).

Researchers are sorting through genetic and molecular biomarkers that could help to forecast the course of an individual tumour. Such predictions may never be 100% accurate, but a little more clarity could go a long way.

2

**Who should be having mammograms, and how often?**

Women face a cacophony of conflicting recommendations regarding screening. Should they be screened every year starting at age 40, every 2 years at age 50 or never? Although over-screening can create false alarms, under-screening could lead to preventable deaths.

Mammograms prevent about 4 cancer deaths for every 1,000 women screened regularly for 20 years. The benefits seem to be especially small for women in their 40s (see page S118).

Radiologists are working to improve the accuracy and clarity of breast imaging, and public health experts are developing evidence-based guidelines that can save lives while minimizing harm.

3

**What are the possibilities, and limits, of immunotherapy?**

Harnessing the immune system to fight breast cancer could be a winning strategy, a major disappointment or both. For certain women, immunotherapy could mean not just a treatment, but also a cure.

With more than 40 clinical trials underway, data continues to roll in (see page S105). So far, immunotherapy, including anti-cancer vaccines, alone or in combination with other therapies, seems to be especially promising for cases of 'triple-negative' breast cancer.

Each case of cancer is different, and so is each immune system. Matching the right treatment with the right patient could prove to be the defining challenge of the future.

4

**What are the risk factors for the disease?**

Knowing which women are susceptible to breast cancer — and why — could illuminate the root causes of the disease and lead to new approaches for prevention and treatment.

Some women are more likely to develop breast cancer because they carry a mutation in a *BRCA* gene (see S116), but even the risk for these women is shaped by exposure to hormones, which in turn are influenced by exercise, weight control and pregnancy.

Researchers are looking beyond *BRCA* for other gene mutations that could help to set breast cancer in motion. The genetic links probably will not be as strong, but finding them could still prove lifesaving.

Chris Woolston is a science writer based in Billings, Montana.



## BREAST CANCER

# Doubtful health benefit of screening from 40 years of age

Philippe Autier

**Refers to** Moss, S. M. *et al.* Effect of mammographic screening from age 40 years on breast cancer mortality in the UK Age trial at 17 years' follow-up: a randomised controlled trial. *Lancet Oncol.* [http://dx.doi.org/10.1016/S1470-2045\(15\)00128-X](http://dx.doi.org/10.1016/S1470-2045(15)00128-X)

**Results of the UK Age trial suggest a significant benefit of annual mammography initiated at 39–41 years of age in preventing breast-cancer deaths occurring before the age of 50 years; however, this approach had no effect on the risk of breast-cancer death occurring before the age of 60 years and leads to prolonged deteriorations in quality of life owing to overdiagnosis.**

Organized programmes of mammographic screening have focused predominately on women aged  $\geq 50$  years; whether screening would be of benefit to younger women, such as those aged 40–50 years, remains unclear. Moss *et al.*<sup>1</sup> have now published long-term results of the UK Age trial, which is the only study that has specifically addressed the population-wide efficacy of mammography screening starting at the age of 40 years.

Moss *et al.*<sup>1,2</sup> randomly assigned women aged 39–41 years into either an intervention group comprising 53,883 women who were invited to undergo annual mammography screening, or a control group of 106,953 women who did not undergo screening from this age; 68% of women invited for annual mammography regularly participated, and these women received a maximum of seven rounds of screening up to the age of 48 years.<sup>1,2</sup> At 50–52 years of age, all of the women were invited to participate in the UK NHS breast-screening programme (mammography every 3 years). Causes of death were recorded up to 20 years after randomization (median follow-up 17.7 years). Moss *et al.*<sup>1</sup> concluded that screening of women aged 40–49 years is associated with an early reduction of breast-cancer mortality based on a statistically significant decrease in the relative risk of this outcome during the first 10 years after initiation of annual screening (RR 0.75, 95% CI 0.58–0.97; Table 1). However, this risk reduction faded away beyond the 10-year time point; at year 20, the risk of breast-cancer death in both groups was not significantly different (RR 0.93, 95% CI 0.80–1.09). What is

the public-health relevance of annual mammography screening from 40 years of age if starting less-frequent screening at an age of 50 years is associated with the same long-term risk of death from breast cancer? This question demands further consideration of the benefits, harms, and costs of earlier screening.

Of note, whether the transient risk reduction observed can be attributed to screening is questionable because of the 'left-to-nature' design of the UK Age trial.<sup>1</sup> According to this design, only women invited to screening as part of the intervention group knew they were enrolled in the clinical trial. Women allocated to the control group were never contacted and, therefore, were unaware that they were participating in a trial. Furthermore, the health professionals involved in the trial knew or could identify the women who were invited to screening, but were unaware of the women comprising the control group. These imbalances might have increased disease awareness and the quality of medical management of women in the intervention group, thus introducing performance biases. Moreover, the knowledge among the health professionals, who were ultimately responsible for completing the death certificates, that certain women attended mammography screening might have led to under-reporting of breast cancer as the underlying cause of death, as suggested in other studies.<sup>3,4</sup>

The possibility that the left-to-nature study design led to an overestimation of risk reductions is supported by two facts. First, the significant reduction in breast-cancer mortality was observed in the first 10 years

when the left-to-nature design was enforced, but not thereafter (Table 1)—when invitations to screening were the same for all women. Most deaths attributable to breast cancer occur several years after diagnosis; therefore, deaths prevented as a result of earlier detection of breast cancers in women aged up to 48 years in the intervention arm would have been expected to translate into a persistently reduced rate of breast-cancer mortality beyond the last round of annual screening, compared with that of the control arm. Thus, the similarity in breast-cancer mortality between these cohorts during the last 10 years of the trial, when considering only the women with cancers diagnosed in the first 10 years (Table 1), indicates that earlier detection of cancers as a result of screening during the intervention period might have been quite limited. Second, using the Nottingham Profile Index that combines tumour size and grade, and the lymph-node status, Moss *et al.*<sup>5</sup> previously predicted a relative risk of breast-cancer death over a 10-year period of 0.90 (95% CI 0.80–1.01) for the women diagnosed with breast cancer during the first 8 years of the trial. This predicted relative risk of breast-cancer mortality is much lower than the relative risk of 0.75 reported in the most-recent publication of the trial,<sup>1</sup> which suggests that the mortality reductions observed in the first 10 years of the trial might be attributable to factors other than screening.

Analysis of the Nelson–Aalen curves of the cumulative breast-cancer mortality for all women with breast cancers diagnosed during the follow-up period of up to 20 years reveals that in the last 10 years, the difference in cumulative mortality between the intervention and the control groups progressively shrank<sup>1</sup>—at year 20, no significant difference was seen.<sup>1</sup> Interestingly, a relatively constant separation of the mortality curves between year 10 and year 20 was observed when only the women with cancers diagnosed during the intervention period were considered.<sup>1</sup> These intriguing trends imply that the mortality associated with breast cancer diagnosed in the last 10 years of the trial was greater in the intervention group than in the control group. To verify this finding, I derived the death rates attributed to breast cancers diagnosed during years 10–20 by subtracting the rates in each of the mortality

**Table 1** | Rates and relative risks of breast-cancer mortality in the UK Age trial\*

Breast-cancer deaths considered	Number of breast-cancer deaths		Women-years		Breast-cancer-death rate per 1,000 women-years		Relative risk of breast cancer death (95% CI)	Rate difference per 1,000 women-years (intervention vs control group)
	Intervention group	Control group	Intervention group	Control group	Intervention group	Control group		
In women diagnosed <10 years after randomization								
Within 10 years	83	219	532,747	1,058,322	0.156	0.207	0.75 (0.58–0.97)	–0.051
Beyond 10 years	99	193	408,221	810,395	0.243	0.238	1.02 (0.80–1.30)	0.005
Up to year 20	182	412	940,969	1,868,717	0.193	0.220	0.88 (0.74–1.04)	–0.027
In women diagnosed ≥10 years after randomization†								
Beyond 10 years	60	103	408,221	810,395	0.147	0.127	1.16 (0.84–1.59)	0.020
In all women diagnosed up to year 20 after randomization								
Up to year 20	242	515	940,969	1,868,717	0.257	0.276	0.93 (0.80–1.09)	–0.018

\*All data come from the publication by Moss *et al.*<sup>1</sup> †Figures calculated based on data from Moss *et al.*<sup>1</sup>

curves; this calculation produced a relative risk of death of 1.16 that was not significant owing to the small numbers of deaths (Table 1). Nevertheless, this result indicates that a persistent excess of deaths in the intervention group during this period contributed to equalizing death rates in the two groups at year 20. These observations echo those of previous studies, which concluded that annual mammographic screening of women in their forties would result in an increase in breast-cancer mortality.<sup>6–8</sup> In particular, periodic exposure of the breasts to X-rays starting at age 39–41 years in the UK Age trial might have increased the risk of breast-cancer death at the later time points.

Overdiagnosis (and thus overtreatment) is another concern associated with use of screening programmes. In the context of screening, overdiagnosis refers to the detection of *in situ* or invasive cancers unlikely to become clinically evident during the patient's lifetime.<sup>9</sup> Moss *et al.*<sup>1</sup> computed rates of overdiagnosis in the UK Age trial by comparing the incidence of such breast neoplasms in the two cohorts, with the investigators suggesting that use of earlier screening was associated with “at worst a small amount of overdiagnosis”.<sup>1</sup> However, the frequency of overdiagnosis in women invited to screening at some point in their life compared with the rate in women never invited to screening remains the more-important question. To examine this question, I extracted data from the article on the cumulative incidence rates of *in situ* and invasive breast cancers and regressed the rates observed during the first 8 years in the control group—when no screening was performed in this cohort. In the absence of mammography screening, the cumulative incidence of breast cancer has been shown to increase linearly with age;<sup>10</sup>

therefore, using the linear-regression slope obtained for the data from the control cohort up to year 8, I computed a 20-year cumulative incidence of breast cancer of 27 per 1,000 women. The quasi-identical cumulative incidence at 20 years reported for both groups in the trial was around 42 per 1,000 women.<sup>1</sup> This difference of approximately 15 cases per 1,000 women suggests a 35% (15/42) frequency of overdiagnosis of *in situ* and invasive breast cancers in both groups. Interestingly, the data on the cumulative incidence of these cancers presented by Moss *et al.*<sup>1</sup> also shows that the level of overdiagnosis is independent of the age at which screening starts, indicating that the reservoir of nonprogressing cancerous lesions that are detectable by mammography is present for long periods of time. Thus, if the rates of overdiagnosis are similar when screening is started at the ages of 39–41 years or 50–52 years, women who participate in screening at an earlier age risk longer periods of deteriorated quality of life (relating to mutilation, including mastectomy, and other adverse effects of treatments, as well as associated detriments in self-image, sexual relationships, psychological distress, and so on) owing to overdiagnosis and overtreatment.

The UK Age trial provides evidence that starting mammography screening at the age of 40 years would provide or no additional health benefit over screening initiated at an age of 50 years, whereas the harmful consequences of such a programme would persist for longer periods. Methods of screening for breast cancer other than mammography need to be developed and evaluated.

University of Strathclyde and International Prevention Research Institute (iPRI) Institute of Global Public Health, iPRI, 95 Cours Lafayette, 69006 Lyon, France.  
philippe.autier@i-pri.org

doi:10.1038/nrclinonc.2015.162

Published online 15 September 2015

#### Competing interests

The author declares no competing interests.

1. Moss, S. M. *et al.* Effect of mammographic screening from age 40 years on breast cancer mortality in the UK Age trial at 17 years' follow-up: a randomised controlled trial. *Lancet Oncol.* [http://dx.doi.org/10.1016/S1470-2045\(15\)00128-X](http://dx.doi.org/10.1016/S1470-2045(15)00128-X) (2015).
2. Moss, S. M. *et al.* Effect of mammographic screening from age 40 years on breast cancer mortality at 10 years' follow-up: a randomised controlled trial. *Lancet* **368**, 2053–2060 (2006).
3. Gøtzsche, P. C. & Jørgensen, K. J. Screening for breast cancer with mammography. *Cochrane Database of Systematic Reviews*, Issue 6. Art. No.: CD001877 <http://dx.doi.org/10.1002/14651858.CD001877.pub5> (2013).
4. Skrabanek, P. Breast cancer screening with mammography. *Lancet* **341**, 1531 (1993).
5. Moss, S., Waller, M., Anderson, T. J., Cuckle, H.; Trial Management Group. Randomised controlled trial of mammographic screening in women from age 40: predicted mortality based on surrogate outcome measures. *Br. J. Cancer* **92**, 955–960 (2005).
6. Berrington de González, A. & Reeves, G. Mammographic screening before age 50 years in the UK: comparison of the radiation risks with the mortality benefits. *Br. J. Cancer* **93**, 590–596 (2005).
7. Mattsson, A., Leitz, W. & Rutqvist, L. E. Radiation risk and mammographic screening of women from 40 to 49 years of age: effect on breast cancer rates and years of life. *Br. J. Cancer* **82**, 220–226 (2000).
8. Beemsterboer, P. M., Warmerdam, P. G., Boer, R. & de Koning, H. J. Radiation risk of mammography related to benefit in screening programmes: a favourable balance? *J. Med. Screen.* **5**, 81–87 (1998).
9. Day, N. E. Overdiagnosis and breast cancer screening. *Breast Cancer Res.* **7**, 228–229 (2005).
10. Miller, A. B. *et al.* Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial. *BMJ* **348**, g366 (2014).

# Precision medicine for metastatic breast cancer—limitations and solutions

Monica Arnedos, Cecile Viciér, Sherene Loi, Celine Lefebvre, Stefan Michiels, Herve Bonnefoi and Fabrice Andre

**Abstract** | The development of precision medicine for the management of metastatic breast cancer is an appealing concept; however, major scientific and logistical challenges hinder its implementation in the clinic. The identification of driver mutational events remains the biggest challenge, because, with the few exceptions of *ER*, *HER2*, *PIK3CA* and *AKT1*, no validated oncogenic drivers of breast cancer exist. The development of bioinformatic tools to help identify driver mutations, together with assessment of pathway activation and dependency should help resolve this issue in the future. The occurrence of secondary resistance, such as *ESR1* mutations, following endocrine therapy poses a further challenge. Ultra-deep sequencing and monitoring of circulating tumour DNA (ctDNA) could permit early detection of the genetic events underlying resistance and inform on combination therapy approaches. Beside these scientific challenges, logistical and operational issues are a major limitation to the development of precision medicine. For example, the low incidence of most candidate genomic alterations hinders randomized trials, as the number of patients to be screened would be too high. We discuss these limitations and the solutions, which include scaling-up the number of patients screened for identifying a genomic alteration, the clustering of genomic alterations into pathways, and the development of personalized medicine trials.

Arnedos, M. et al. *Nat. Rev. Clin. Oncol.* advance online publication 21 July 2015; doi:10.1038/nrclinonc.2015.123

## Introduction

The discovery of the oestrogen receptor (*ER*)<sup>1</sup> and human epidermal growth factor receptor-2 (*HER2*)<sup>2</sup> as therapeutic targets in patients with breast cancer has enabled treatment success in terms of patient outcomes with *ER* or *HER2*-blocking therapies,<sup>3,4</sup> and set the stage for the development of stratified medicine. Furthermore, progress in cancer genomics research over the past few decades has reinforced the notion that cancer is driven by various genomics alterations.<sup>5</sup> As a result of different international initiatives such as The Cancer Genome Atlas (TCGA) or the International Cancer Genome Consortium (ICGC), the use of next-generation sequencing (NGS) has helped define the genomic landscape of early stage breast cancer.<sup>6</sup> These studies have revealed the high level of tumour heterogeneity for each breast tumour that consists of several molecular subsets, which are driven by distinct molecular alterations, indicating that tumours could be treated according to their individual molecular landscape. Despite the exciting potential for personalized medicine, *ER* and *HER2* are currently the only targetable molecular alterations with confirmed predictive and prognostic value.<sup>3,4</sup> Other targeted therapies, such as *mTOR* and *CDK4/6* inhibitors, have been approved on the basis of

their efficacy in subgroup populations, but no predictive biomarkers have been found.<sup>7,8</sup> In this Review, we discuss the potential applications of genomics to improve the management of metastatic breast cancer (MBC), and consider the challenges that precision medicine must overcome before it can be widely implemented in the clinic—most notably, those challenges that relate to the remarkable cellular complexity of this type of cancer.

## Genomic landscape of breast cancer

Analysis of the molecular features of early stage breast cancer using NGS has led to the description of the genomic landscape of this disease.<sup>6</sup> This research has confirmed that *TP53* and *PIK3CA* mutations are the most frequent genomic alterations overall in all intrinsic subtypes (28% for both genes). Amplifications in *ERBB2*, *FGFR1* and *CCND1* follow in frequency, being observed in 10–20% of all breast cancer subtypes. Additional alterations are less frequent, but could be highly clinically relevant, including *PTEN* mutations and deletions, and *AKT1*, *RB1*, *BRCA1* or *BRCA2* mutations. Sequencing analyses have uncovered mutations in other genes of interest that might have some clinical relevance in breast cancer, including *KRAS*, *APC*, *NF1*, *SKT11*, *MAP2K4*, *MAP3K1* and *AKT2*.<sup>6</sup> A similar genomic study focused on patients with triple-negative breast cancer (TNBC) revealed a more heterogeneous molecular profile, with some tumours having just a few molecular alterations whereas others harboured hundreds of alterations.<sup>9</sup>

Department of Medical Oncology (M.A., F.A.), INSERM Unit U981 (C.V., C.L.), Department of Biostatistics and Epidemiology (S.M.), Gustave Roussy and Université Paris Sud, 94800 Villejuif, France. Division of Research and Cancer Medicine, Peter MacCallum Cancer Centre, University of Melbourne, East Melbourne, VIC 3002, Australia (S.L.). Department of Medical Oncology and INSERM U916, Institut Bergonié, 229 Cours de l'Argonne, 33000 Bordeaux, France (H.B.).

Correspondence to: F.A. [fabrice@igrc.fr](mailto:fabrice@igrc.fr)

## Competing interests

F.A. receives honourarium and has a research contract with AstraZeneca and Novartis. M.A. receives honourarium from Novartis. The other authors declare no competing interests.



## Key points

- Recent research data defining the genomic landscape of breast cancer have reinforced the notion that this disease is driven by genomic alterations
- So far very few gene drivers validated in breast cancer have been identified, including *ESR1*, *ERBB2*, *PIK3CA* and *ATK1*
- Identification of drivers, characterization of resistant clones, identification of DNA repair defects and mechanisms of immune suppression are potential uses of genomics to personalize medicine
- The development of precision medicine for the treatment of breast cancer has several major challenges that include low frequency of targetable molecular alterations, feasibility of high-throughput technologies and availability of targeted therapy

**Table 1** | Targetable genomic alterations in breast cancer

Gene	Alteration	Frequency (%)	Candidate drug	Level of evidence for the target
<b>Growth factor receptors</b>				
<i>ERBB2</i>	Amplifications Mutations	>10	HER2 inhibitor	1 3
<i>FGF3</i>	Amplifications	5–10	FGFR inhibitor	4
<i>FGFR1</i>	Amplifications	5–10	FGFR inhibitor	2
<i>FGFR2</i>	Amplifications	1–5	FGFR inhibitor	2
<i>IGF1R</i>	Amplifications	1–5	IGFR inhibitor	4
<i>EGFR</i>	Amplifications	1–5	EGFR inhibitor	2
<b>PI3K/AKT/mTOR</b>				
<i>PIK3CA</i>	Amplifications Mutations	>10	PI3K inhibitor	1–2
<i>PIK3R1</i>	Mutations	1–5	Not known	4
<i>PTEN</i>	Mutations Deletions	5–10	AKT inhibitor	3
<i>AKT1</i>	Amplifications Mutations	1–5	AKT inhibitor	2
<i>AKT2</i>	Amplifications	1–5	AKT inhibitor	2
<i>AKT3</i>	Amplifications	1–5	AKT inhibitor	4
<i>INPP4B</i>	Deletions	1–5	AKT inhibitor	NA
<b>MEK pathway</b>				
<i>NF1</i>	Mutations	1–5	MEK inhibitor	2c
<i>KRAS</i>	Amplifications	1–5	MEK inhibitor	2c
<i>BRAF</i>	Amplifications	1–5	MEK inhibitor	2c
<b>JNK pathway</b>				
<i>MAP2K4</i>	Mutations Deletions	5–10	Not known	NA
<i>MAP3K1</i>	Mutations Deletions	5–10	Not known	NA
<i>GPS2</i>	Mutations	1–5	Not known	NA
<b>Cell cycle</b>				
<i>CCND1</i>	Amplifications	>10	CDK4 inhibitor	4
<i>CDKN2A</i>	Deletions	5	Not known	NA
<i>CDKN1B</i>	Alterations	1–5	Not known	NA
<i>CDK4</i>	Amplifications	1–5	CDK4 inhibitor	4
<i>Rb</i>	Mutations Deletions	5–10	Resistance to CDK4 inhibitor	3
<b>DNA repair</b>				
<i>BRCA1</i>	Mutations Deletions	1–5	PARP inhibitor	1

Interestingly, a comparison of the RNA sequencing with the genomes/exomes in this study revealed that only 36% of validated somatic single nucleotide variations were observed in a transcriptome sequence, which poses a question regarding the use of NGS alone to identify potential drivers of breast cancer. Finally, proteomic analyses in relation to the TCGA dataset in early stage breast cancer have indicated that three pathways are predominantly activated in all subtypes: PI3K/AKT/mTOR, p53 and CCND1/CDK4/Rb.<sup>6</sup> A large proportion of the molecular alterations identified can be targeted by new therapeutics; a non-exhaustive list of candidate targetable genomic aberrations is shown in Table 1.

Molecular analyses of metastatic breast cancers suggest that breast cancer can evolve over time, losing and/or gaining new alterations. As an illustration, discordances in ER, progesterone receptor (PgR), and HER2 status between the primary tumour and the metastasis have been reported in around 16%, 40%, and 10% of patients, respectively,<sup>10</sup> with loss of HER2 expression reported in around 24% of patients.<sup>11</sup> Several studies have investigated the genomic landscape of breast cancer from early to metastatic disease, identifying *ESR1* mutations as one of the genomic alterations that mediates resistance to aromatase inhibitors (AI).<sup>12,13</sup> *ESR1* mutations occur in 10–30% of ER-positive MBC that are resistant to AIs, and lead to ligand-independent activation of the ER. Whether *ESR1* mutations are present as a minor subclone when the cancer arises or are acquired during the treatment is still unknown; however, the percentage of these mutations increases under the selective pressure of therapy and are able to drive resistance. So far, few large comprehensive analyses have reported the genomic landscape of metastatic breast cancer. One of these studies performed whole-exome sequencing (WES) of 93 biopsy samples from patients with MBC, and reported, in addition to the already known molecular aberrations in the early setting, a higher incidence of *TSC1/TSC2* mutations.<sup>14</sup> In the study by Toy *et al.*,<sup>12</sup> *TP53*, *PIK3CA* and *GATA3* were the most frequently identified molecular alterations.<sup>12</sup> In addition to the previously mentioned mutations in *ESR1*, these investigators have identified mutations in *RPTOR* and *ERBB3* at much higher rates than those reported in the TCGA. In the latest interim analysis of the currently ongoing MOSCATO-01 trial,<sup>15</sup> a molecular screening of 700 patients with advanced-stage cancer, including around 70 with metastatic breast cancer, alterations of the PTEN/PI3K/AKT and FGFR/FGF pathways were the two most frequently detected actionable pathways observed across all tumour types. The now proven evolution of breast cancer over time constitutes the basis for assessing metastases to direct therapy in trials of precision medicine. Other potential mechanisms of acquired resistance have been reported in other disease subtypes, including reversal of *BRCA1/2* mutations as a mechanism of secondary resistance to PARP inhibitors<sup>16</sup> or, in the HER2-positive segment, the acquisition of *ERBB2* mutations.<sup>17</sup> These mechanisms need to be confirmed, and potential new aberrations have yet to be identified.

**Table 1** (cont.) | Targetable genomic alterations in breast cancer

Gene	Alteration	Frequency (%)	Candidate drug	Level of evidence for the target
<b>DNA repair (cont.)</b>				
<i>BRCA2</i>	Mutations Deletions	1–5	PARP inhibitor	1
<i>ATM</i>	Mutations	1–5	PARP inhibitor	3
<i>ATR</i>	Mutations	1–5	PARP inhibitor	3
<i>MDM2</i>	Amplifications	1–5	MDM2 inhibitor	4
<i>P53</i>	Mutations	>10	Not known	NA
<b>ER signalling</b>				
<i>ESR1</i>	Mutations Amplifications Translocations	>10% in metastatic ER+ MBC resistant to endocrine therapy	Not known	2
<i>GATA3</i>	Mutations	5–10	Endocrine therapy	3
<i>FoxA1</i>	Mutations	1–5	Endocrine therapy	3
<b>Epigenetics</b>				
<i>KMT2C</i>	Mutations	5–10	Drug targeting epigenetics	4
<i>KMT2B</i>	Mutations	1–5	Drug targeting epigenetics	4
<i>KDM6A</i>	Mutations	1–5	Drug targeting epigenetics	4
<i>SETD2</i>	Mutations	1–5	Drug targeting epigenetics	4
<b>Others</b>				
<i>NOTCH3</i>	Amplifications	1–5	NOTCH inhibitor	4

Abbreviations: HDAC, histone deacetylases; NA, not available.

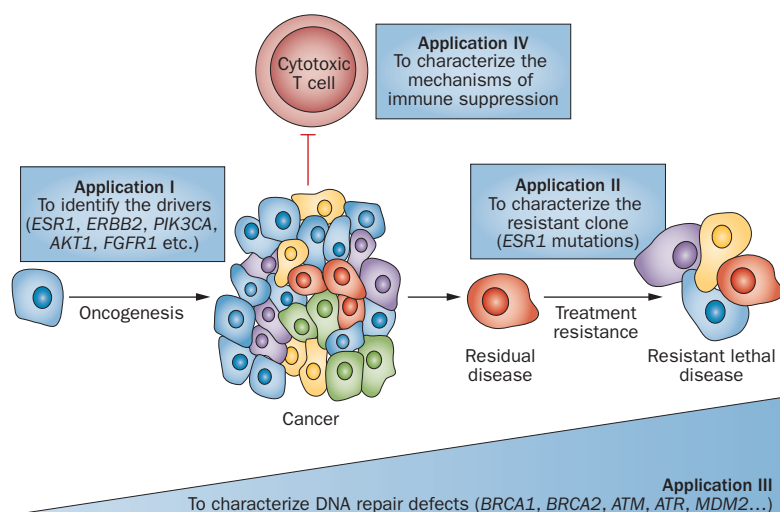
### Tools for precision medicine

The detection of genomic alterations in patients with breast cancer can help to tailor therapy. Over time, several molecular tools have been developed to serve this type of personalized therapeutic approach in breast cancer. Historically, immunohistochemistry was used to stratify patients with breast cancer according to the presence of certain biomarkers, and is currently used to determine the expression of ER and HER2.<sup>18,19</sup> Fluorescence *in situ* hybridization (FISH) analysis was then developed as a method to quantify copy number. Currently, identification of *ERBB2* amplifications is the main standard application of this technique; in the future, it could also be used for the quantification of other amplifications such as *FGFR1* or *CCND1*. Several multigene assays have also been developed to quantify gene-expression using either RT-PCR (Oncotype DX®),<sup>20</sup> DNA array (Mammaprint®),<sup>21</sup> or NanoString technologies (Prosigna™).<sup>22</sup> These assays are currently used in early stage breast cancer to stratify patients according to their risk of relapse, but these technologies could be applied to other applications in the metastatic setting. For example, gene-expression arrays can accurately quantify gene expression and, therefore, determine which pathways might be activated, such as mTOR<sup>23</sup> or MEK,<sup>24</sup> in order to target the most relevant pathways in a particular tumour. More recently, comparative genomic hybridization (CGH) arrays and single nucleotide

polymorphism (SNP) analyses have been used to identify targetable genomic alterations in breast cancer.<sup>25</sup> These technologies quantify DNA copy number in the whole genome, and enable the detection of gene loss and gene amplification, and their application in clinical practice is feasible in around 70% of patients.<sup>25,26</sup> As mentioned before, the main genomic alterations in breast cancer include *FGFR1*, *CCND1*, *ERBB2*, *EGFR*, *MDM2*, *AKT3* amplifications, and *PTEN* deletions.<sup>25</sup> NGS is currently used to detect mutations (and in some instances, copy-number changes) and has several potential applications in the clinic. In most trials NGS is used to detect clonally dominant mutations in a multigene panel, although, ultradeep sequencing can be used to detect minor sub-clonal alterations that would potentially mediate further resistance to targeted therapy. Some centres are starting to use WES in the clinical setting.<sup>27</sup> In the context of breast cancer, WES could be particularly appealing to detect genome scars mediated by DNA repair defects, which could be used as a predictive biomarker for the use of PARP inhibitors.<sup>28</sup> Other applications include the detection of driver alterations and neoantigens. New tools are also being developed that could be complementary to the ones mentioned previously. For example, RNA-seq could be used to quantify RNA expression and detect structural changes, such as mutations and translocations. The detection of ctDNA is also becoming increasingly popular, and has been shown to be feasible in providing a readout of the tumour genomic landscape in patients with breast cancer.<sup>29,30</sup> This technology is being developed as a noninvasive alternative to biopsies to monitor drug efficacy and to detect genomic alterations involved in resistance. Finally, technologies centred on nucleic acid detection will not be sufficient to obtain a comprehensive molecular profile of tumours, as protein expression patterns also need to be considered. Several technologies are being developed, and include TheraLink assay, which measures phosphoprotein levels in tumour samples to identify pathways that are activated. One could, therefore, hypothesize that quantifying the levels of HER2, ER, mTOR and CDK4 expression could be useful to identify the subset of patients that could benefit from approved targeted therapies.

### Genomics and lethal breast cancers

Within the context of metastatic breast cancer, four potential applications of genomics to improve patient outcomes are currently available (Figure 1). The first one is the identification of oncogenic drivers. A genomic driver could be defined as the molecular alteration responsible for cancer progression, thus, targeting this gene is expected to have a therapeutic effect, namely circumvention of oncogene addiction.<sup>31</sup> ER and HER2 are the historical drivers of breast cancer. Other potential drivers, which have been identified by high-throughput technologies, are potential targets of genomic-driven drug development.<sup>6,14,32</sup> Most of these mutated genes, with the exception of *PIK3CA*, are present at low frequencies (<10%) and have yet to be validated. On the one hand,



**Figure 1** | Possible applications of genomics to personalize therapy of MBC. A first application is the identification of new gene drivers for MBC, aside from *ER*, *ERBB2*, *PIK3CA* and *AKT1*. Further applications include the characterization of the molecular alterations specific for the resistant clones (application II), such as the presence of *ESR1* mutations in patients previously treated with an aromatase inhibitor; the identification of DNA repair defects that might influence treatment decision (application III) and the characterization of mechanisms of immune suppression developed by the tumour cell (application IV). Abbreviation: MBC, metastatic breast cancer.

several genomic alterations have already been associated with an objective response when targeted clinically. These include *PIK3CA* mutations,<sup>33</sup> *FGFR1* amplification,<sup>34,35</sup> *AKT1* mutations,<sup>25</sup> and *EGFR* amplifications.<sup>25</sup> In addition, *ERBB2* mutations present a very strong preclinical rationale.<sup>36</sup> On the other hand, some genomic alterations have been reported as candidate drivers, but have not been found to be predictive for the efficacy of targeted therapies. This was the case for *CCND1* amplifications and the CDK4/6 inhibitor palbociclib.<sup>37</sup> This finding indicates that some drivers of cancer progression could be detected through analysis at the RNA/protein level, and not only at the DNA level. In fact, ER expression is the most important driving force of cancer progression and, notably, no DNA alterations in *ESR1* are detected in most patients.<sup>12</sup> The same principle could apply to mTOR and CDK4 inhibition, as these pathways are activated in most tumour samples, even in the absence of any pathway specific genomic alterations.<sup>38,39</sup> Thus, there is a need to develop molecular tools that measure pathway activation, rather than gene amplifications or mutations, either through RNA or protein expression.

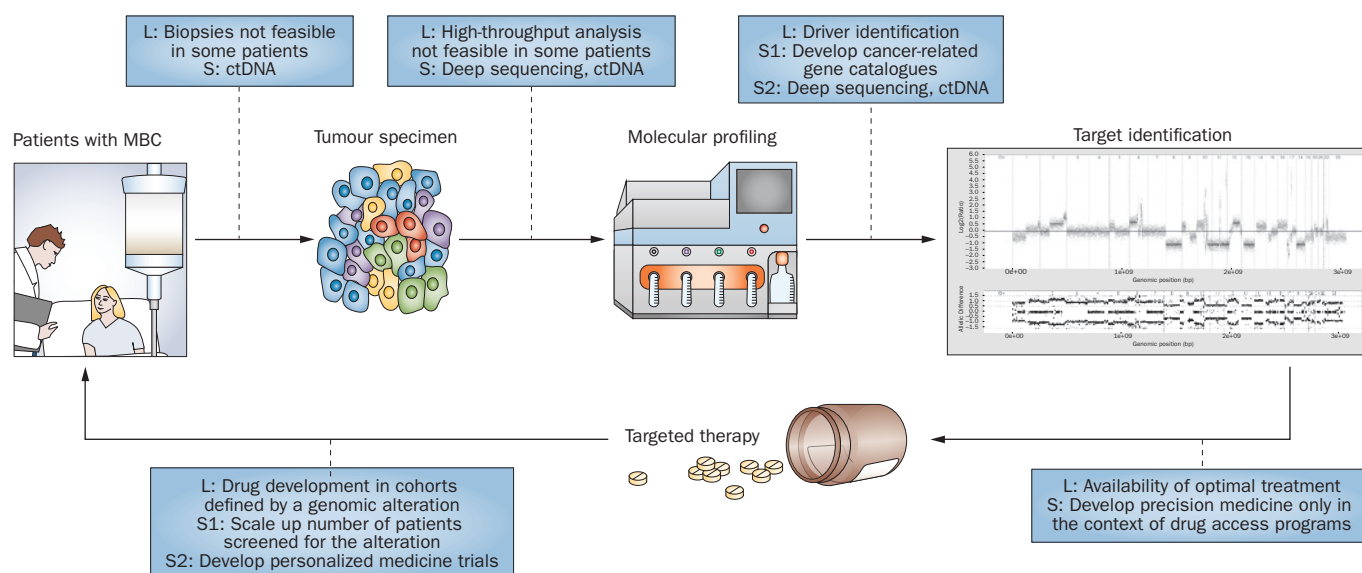
The second possible application of genomics in advanced-stage breast cancer is the identification of genomic alterations responsible for secondary resistance. As discussed previously, *ESR1* mutations are observed with frequencies of up to 10–30% in tumours previously treated with AIs.<sup>12,13</sup> Moreover, preclinical studies have demonstrated that the presence of these mutations renders the cancer cell insensitive to AIs, but partly sensitive to endocrine therapies that target ERs directly, as these mutations, which mostly affect the ligand-binding domain, induce the production of a constitutively active receptor with a conformational structure that limits

the binding of agents to the receptor.<sup>12</sup> Comparison of TCGA and WES results have shown that these molecular alterations are not present or are present in very low levels in the pretreatment samples, even when detected using ultra-deep sequencing.<sup>12,13</sup> In April 2015, the ER-degrading agent GDC-0810 has demonstrated anti-tumour activity in patients with ER-positive tumours, even when the tumours harboured an *ESR1* mutation.<sup>40</sup> Thus, detecting *ESR1* mutations, or any other secondary mutations, in patients with advanced-stage breast cancer could enable treatment with a specific targeted therapy. Of note, other molecular alterations that could be involved in secondary endocrine resistance include *TSC1/2* mutations that could be targeted using an mTOR inhibitor.<sup>14</sup> Similarly, other genomic alterations that might be involved in resistance to targeted therapies, such as *PTEN* mutations and deletions, have been reported to mediate resistance to  $\alpha$ -selective PI3K inhibitors.<sup>41</sup> Secondary mutational events could be determined by performing ctDNA analysis and ultradeep sequencing; analysis of ctDNA might enable these genomic alterations to be detected early in disease evolution (even before relapse) and ultradeep sequencing could enable detection of these mutations as soon as the tumour becomes detectable. Nevertheless, it must be acknowledged that no evidence exists in breast cancer that ultradeep sequencing will lead to the detection of minor subclonal alterations, such as the previously mentioned *ESR1* or *PTEN* mutations. Whether detecting these genomic alterations early during the disease course will lead to improvement in outcomes remains a major clinical question in the field of precision medicine.

A third application of genomics in lethal breast cancer is the identification of DNA-repair defects, mutational processes and defects in the DNA duplication mechanism. Indeed, the use of high-throughput WES could identify the defective DNA repair pathway in each patient. For example, a study using WES to analyse 21 primary breast tumours and matched normal DNA samples from the same individuals<sup>42</sup> identified five biologically distinct mutational signatures present in these 21 cancers (named A–E), characterized by a different expression profile, and each signature contributing to the cancer development in all 21 patients, each to a different extent.<sup>42</sup> This study also identified the apolipoprotein B mRNA-editing enzyme catalytic polypeptide (APOBEC) family of cytidine deaminases as having a role in generating particular genome-wide mutational signatures, including a signature of localized hypermutation (kataegis).

In fact, the APOBEC enzymes, which have evolved as key players in natural and adaptive immunity, have been proposed to contribute to cancer development and clonal evolution of cancer by inducing collateral genomic damage owing to their DNA deaminase activity.<sup>42</sup> Data also suggest that high APOBEC activity is a source of *PIK3Ca* mutations in the helical domain, whereas in tumours with low APOBEC activity, *PIK3Ca* is equally likely to be mutated at the kinase domain hot spot and the helical domain.<sup>43</sup>





**Figure 2** | Limitations and solutions for precision medicine in MBC. Several limitations and possible solutions to the development of a precision medicine approach for the management of MBC are depicted in this figure. Some limitations are related to the tumour specimen and its molecular profiling; while others are related to the identification of driver genomic alterations to guide targeted therapy. Even when a potential driver has been identified, several challenges arise in selection of the targeted therapy, including availability of optimal therapy and drug development in cohorts defined by a genomic alteration. Possible solutions are also reported. Abbreviations: ctDNA, circulating tumour DNA; L, limitation; MBC, metastatic breast cancer; S, solution.

DNA-repair deficiencies can also be detected using other technologies. For example, the use of SNP arrays could help to detect those tumours with homologous recombination deficiencies that might be sensitive to PARP inhibition.<sup>44</sup> Sequencing of candidate genes could help to identify DNA repair defects in the tumour and lead to development of individualized treatment strategies. As an illustration, PARP inhibitors have been shown to be effective specifically in patients with *BRCA1/2* mutations.<sup>45</sup> Other genes of interest in this category include *ATM* and *ATR*.<sup>46,47</sup> Finally, detection of *TOP2A* amplifications could be useful to determine which patients should receive anthracyclines.<sup>48</sup>

The fourth potential application of high-throughput technologies in the treatment of advanced-stage breast cancer is the identification of mechanisms of immune escape at the individual level. Anti-PD1 treatments have shown antitumour activity as monotherapy in TNBC, with an objective response of 18.5%.<sup>49</sup> Nevertheless, tools that might predict efficacy of such drugs are currently lacking, and there is a need to find targets to develop other immunotherapeutics. Genomics can potentially be used to evaluate five different aspects of the immune system for the purpose of precision medicine. First, it has been suggested from studies of patients with melanoma and lung cancer that the detection of neoantigens and mutational load could predict efficacy of immune checkpoint modulators.<sup>50</sup> The applicability of such a concept to breast cancer is still unknown. Second, genomics could be used to quantify expression of ligands of immune checkpoint proteins. Third, sequencing and gene-expression analysis could detect those cancer cells that are resistant to cytotoxic T-cell lymphocytes (CTL). For example, some

cancer cells present either a *TAP1* mutation or they do not express MHC class I; both of them confer resistance to CTL.<sup>51</sup> Fourth, genomics could be used to quantify and qualify the local immune system. For example, infiltration by CD8<sup>+</sup> T cells could predict efficacy of anti-PD1 therapy;<sup>52</sup> thus, gene-expression analysis could quantify such infiltration and determine the level of Th<sub>1</sub> immune response for a given tumour.<sup>53</sup> Finally, genomics could be used to detect genetic polymorphisms that are associated with immune defects. Of note, some breast cancers harbour SNPs in *TLR4* and *P2RX7* that can be associated with immune suppression and lack of response to immunotherapeutics or chemotherapy.<sup>54</sup> All these approaches could have a potential clinical application, and their use in the case of neoantigens would seem particularly promising.

### Limitations of precision medicine

Despite the continuing advances in high-throughput technologies and their multiple applications in cancer, there are several challenges that need to be resolved before they can be applied for personalized medicine in metastatic breast cancer (Figure 2). The most important limitations and some potential solutions are described below. These challenges can be divided into 'scientific' challenges versus 'logistical and operational' challenges (Table 2).

### Driver identification

Most of the genomic-driven trials failed to report high levels of antitumour activity of therapy in patients with metastatic breast cancers.<sup>25,34,55</sup> The biggest challenge to develop precision medicine in breast cancer will be to

**Table 2** | Limitations of precision medicine for breast cancer, and possible solutions

Limitations	Evidence	Possible solutions
<b>Logistical and operational challenges</b>		
Completion of trials testing drugs in genomic segments is challenging	Genomic alterations are rare, yet randomized clinical trials are still required for approval	Scale-up molecular screening and cluster genomic alterations into pathways Move to personalized medicine Develop fast-track approval based on comparison with historical controls
Genomic results cannot be provided for a substantial number of patients	Biopsy is not feasible in all patients, low percentage of cancer cells in some patients	Collection of ctDNA and analysis using in-depth NGS
Number of patients with an identified oncogenic driver is low	Number of genes screened is too low Known DNA alterations do not explain cancer progression in a large number of patients	Use multiplex assays Use RNA and protein-based assays to identify drivers
Drug access is limited	Drug development is limited to small number of locations and patients	Develop programs for drug access Limit the panel genes to those for which drugs are available
Genomic tests are too expensive and therefore not affordable by the majority	Genomic tests for clinical use are run by private companies	Run genomic tests in academic centres to allow access to innovation for all patients
<b>Scientific challenges</b>		
Response rates are low	Failure to identify oncogenic drivers Multiple pathways activated Intratour heterogeneity	Improve tools to identify drivers Develop protein and RNA based predictors Develop treatment or drug combinations Exclude patients with high intratour heterogeneity?
Secondary resistance occurs	Additional genomic alterations under treatment pressure Tumour adaptation using protein networks Tumour evolution	ctDNA and ultradeep sequencing to detect resistance early on and adapt therapy Identify feedback loops and combine therapies Combine with drugs that targets host (immunotherapeutics) and driving forces of tumour evolution (DNA repair deficiency)

Abbreviations: ctDNA, circulating tumour DNA; NGS, next-generation sequencing.

increase the magnitude of the antitumour effects associated with such an approach; therefore, it is paramount to identify and target the actionable genomic driver events and differentiate them from passenger events. This is a substantial issue because (with the exception of ER, *ERBB2* and *PIK3CA*) none of the genomic alterations reported in any of the sequencing projects<sup>6,32</sup> have been validated as a target in the preclinical or clinical settings. Driver alterations are expected to belong to the ‘cancer-related’ gene catalogue, proposed by the Broad Institute in 2013.<sup>56</sup> Also, these alterations are expected to be clonally dominant and to lead to alterations in protein function (including activating mutations, biallelic loss or high level amplification in a narrow amplicon). The definition of cancer-related genes together with the listed driver characteristics are an ongoing effort led by bioinformaticians to produce an algorithm that will aid in the identification of these *bona fide* drivers. Most of the bioinformatic methods to identify driver genes exploit the recurrence of alterations to identify those genes that, when altered, might confer a survival advantage to the tumour cells. In particular, methods such as Gistic2<sup>57</sup> or RAE,<sup>58</sup> enable detection of loci with recurrent focal copy-number amplifications or losses, and identification of functional copy-number alterations. More recently, with the growing number of large-scale sequencing projects and the associated catalogues of somatic mutations, computational biologists have focused on the development of methods for

the identification of highly mutated genes. These methods aim to identify those genes that demonstrate features of positive selection by taking into account the mutational load of the gene and comparing it to its background mutational rate,<sup>56,59–61</sup> together with additional features such as the gene-expression level and replication rate<sup>56</sup> or the localization of the mutations and their predicted effects on the activity of the protein.<sup>61</sup> Importantly, these methods will also help categorize the somatic alterations as gain or loss of function, hence, determining the role of the gene as an oncogene or a tumour suppressor, eventually leading to development of targeted therapies.<sup>62</sup> Finally, while these methods mainly exploit DNA alterations, it is important to take into account RNA expression levels (determined for example by gene-expression array) that have been successfully used to identify points of additions of the tumour cells and that should be integrated with DNA alterations for better predictions of targetable pathways in cancer.<sup>63</sup>

The ER is the most validated target to date in breast cancer, where altered protein expression is the key feature, any new target should be defined by protein overexpression and/or pathway activation, rather than DNA alterations alone. Several arguments support the hypothesis that pathway and protein activation could be as important as DNA mutations in breast cancer. First, a significant proportion (around 50%) of patients with breast cancer do not present targetable molecular

alterations in cancer-related genes.<sup>25</sup> Second, activation of pathways and proteins assessed by gene expression<sup>23</sup> or protein phosphorylation<sup>64</sup> is predictive of the efficacy of targeted therapy. Finally, the presence of a target does not necessarily lead to pathway activation. For example, patients with *ERBB2* amplification present high levels of intertumour heterogeneity regarding AKT phosphorylation. Despite this heterogeneity, there is a correlation between *ERBB2* amplification and AKT phosphorylation, but this is not always present in all cases of HER2-positive disease.<sup>65</sup> Overall, with a few exceptions, it is not possible to define a driver at the individual level in breast cancer, although some efforts are ongoing to provide bioinformatics tools that will overcome this issue. A strong rationale exists to support assessing pathway activation, rather than only DNA alteration, to identify potential targetable drivers in breast cancer.

### Targeting molecular alterations: the right drug

The failure to deliver personalized medicine is often associated with the lack of highly bioactive and specific drugs. For example, *PIK3CA* mutations occur in approximately 25% of breast cancers and have been reported as drivers of this disease.<sup>32</sup> Nevertheless, the use of non-selective PI3K inhibitors led to modest response rates (4%) in early clinical trials when administered as monotherapy.<sup>66</sup> Second-generation,  $\alpha$ -selective, PI3K inhibitors are more specific and produce better inhibition *in vivo* in animal models.<sup>67</sup> In fact, preliminary results with these new inhibitors indicate partial responses of around 6% in patients with *PIK3CA* mutant breast cancers, whereas no responses had been observed so far in patients with *PIK3CA* wild-type tumours.<sup>68</sup> Thus, it is crucial to develop drugs that hit the target with the highest specificity and bioactivity.

### Towards combination therapies

Targeting a mutational driver should lead to an antitumour response. Nevertheless, intratumour heterogeneity could explain some primary resistance to therapy. The presence of two or more alterations in cancer-related genes has been associated with resistance to targeted therapies *in vitro* and in clinical trials.<sup>69</sup> Interestingly, Stephens *et al.*<sup>6</sup> reported that 67% of the breast cancer samples analysed had two or more genomic alterations. There are several illustrations of such an issue in breast cancer. For example, the coexistence of *PIK3CA* mutations and *ERBB2* amplification has been associated with resistance to the HER2-targeting drugs trastuzumab and lapatinib,<sup>70</sup> providing the rationale to combine PI3K and HER2 inhibitors in different clinical trials. Interestingly, this resistance to anti-HER2 therapy owing to the presence of *PIK3CA* mutations have not been observed when dual HER2 blockade was achieved through the use of two monoclonal antibodies,<sup>71</sup> or in trials performed in the early setting, where the presence of *PIK3CA* mutations is already evident.<sup>72</sup> Moreover, in the phase III BOLERO2 trial, which evaluated the administration of the mTOR inhibitor everolimus in combination with the aromatase inhibitor exemestane in patients

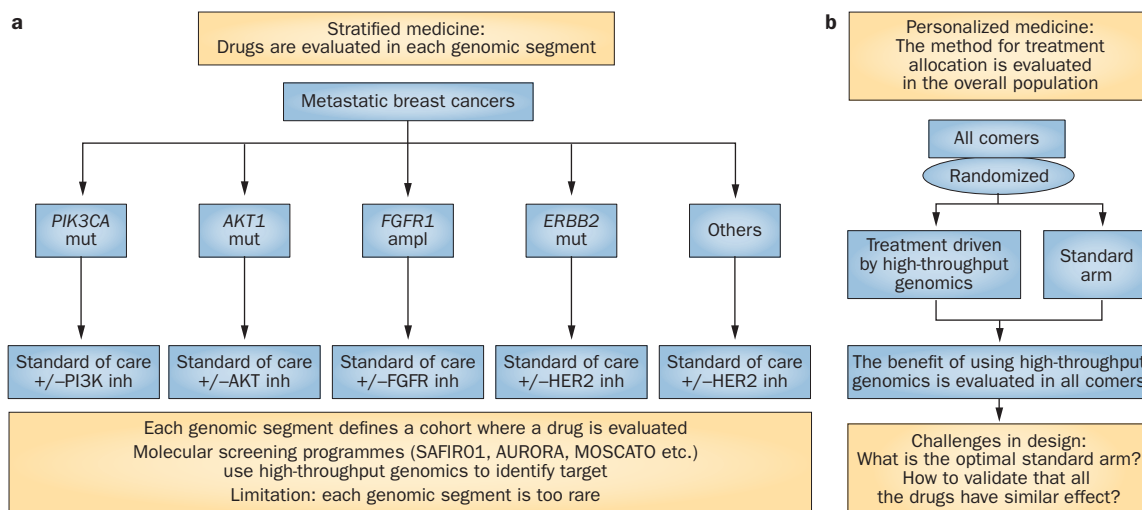
with advanced-stage ER-positive/HER2-negative breast cancer, the presence of multiple alterations in different pathways (that is, *FGFR1/2*, *CCDN1*, *PTEN* and *PIK3CA*) translated into resistance to mTOR inhibition.<sup>7</sup>

The identification of a driver is crucial for the success of targeted therapies, although, it is worth noting that secondary resistance will occur in the vast majority of patients presenting with advanced disease. This resistance is a result of tumour adaptation at the DNA or protein level under the selective pressure of therapy.<sup>73,74</sup> The acquisition of *ESR1* mutations during endocrine therapy is currently the best model of genomic evolution in response to targeted therapy.<sup>12,13</sup> Development of secondary resistance with the emergence of resistant clones could be a result of the high level of intratumour heterogeneity. Several reports emphasize the importance of intratumour heterogeneity and clonal evolution in breast cancer patients. Cancers evolve by a process of clonal expansion, genetic diversification and clonal selection with complex dynamics and with highly variable patterns of genetic diversity and resultant clonal architecture. Therapeutic intervention might decimate cancer clones, but might also promote the expansion of resistant variants.<sup>75,76</sup> From a clinical standpoint, early detection of resistance is crucial to optimizing therapy, and the use of ultradeep sequencing in multiple regions of a biopsy, together with monitoring of tumour evolution using ctDNA,<sup>29</sup> could provide this remit.<sup>77</sup> Moreover, some studies have reported heterogeneity in terms of different metastatic sites exhibiting different genomic alterations,<sup>78</sup> and how to handle this level of heterogeneity is still unclear. One could argue that trials investigating precision medicine should assess genomic alterations only in metastatic sites presenting evidence of progressive disease, but not in those sites that are under control. This approach is being evaluated in an ongoing clinical trial (MATCH-R), which will prospectively investigate the evolution of clonal architecture in tumours from patients that experience disease progression after an initial benefit from a targeted agent. These patients would undergo a biopsy of the progressing lesions in order to identify the mechanisms involved in resistance, and mouse avatars will be generated in some cases to identify the best ad-hoc treatment. At the very least, the presence of multiple drivers in different progressing metastatic sites should lead to the use of combination therapy.

A second mechanism of tumour adaptation consists of the activation of alternative protein networks that will bypass target inhibition. This mechanism of resistance has been documented in patients treated with mTOR inhibitors, whereas a feedback loop lead by mTORC2, has been observed that results in AKT activation owing to growth factor receptor phosphorylation (including IGF-1R),<sup>79</sup> providing the rationale to combine these agents with either IGF-1R<sup>80</sup> or PI3K<sup>81</sup> inhibitors.

Ultimately, one could argue that secondary resistance can only be avoided by developing therapies that could lead to cancer cell eradication. As an illustration, the combination of targeted therapies and the stimulation of





**Figure 3** | Development of MBC precision medicine. There are two main approaches to the implementation of personalized medicine in the management of advanced-stage breast cancer. **a** | A stratified medicine approach: a specific molecular alteration is determined in a tumour sample (pre-screening). Only those tumours harbouring the genomic aberration are treated with the drug targeting that specific alteration. In the case of negativity for this molecular aberration, potentially other alterations might be subsequently determined to inform selection of other targeted therapies. These targets might also be identified within the context of molecular screening programmes that use high-throughput technologies to identify several molecular alterations in parallel. The main limitation of this approach is the low frequency at which the majority of these genomic aberrations are present. **b** | A personalized medicine approach: the goal is to develop trials showing that a drug is effective in a cohort of patients defined by a genomic alteration, because this approach assesses if treatment individualization based on high-throughput genomics improves outcome in all comers as compared with standard therapy. The main challenge for this approach is determining the standard arm (such as targeted therapy but unselected or standard of care). Abbreviations: inh, inhibitors; MBC, metastatic breast cancer.

immune checkpoints could enable eradication of tumour cells.<sup>73</sup> This concept is current being evaluated in a trial testing the combination of trastuzumab and pembrolizumab, an anti-PD1 agent, in HER2-overexpressing breast cancers.<sup>82</sup>

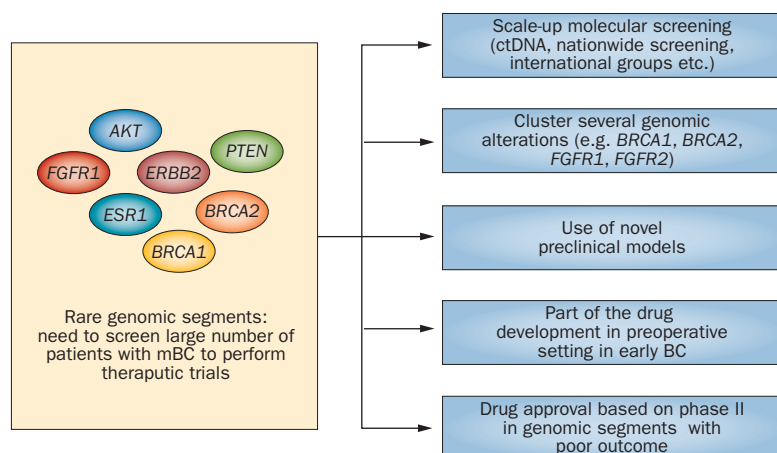
### Logistical and operational challenges

Precision medicine includes two different approaches, namely, stratified medicine and personalized medicine (Figure 3). Stratified medicine (Figure 3a) consists of testing a drug in a population defined by a specific molecular alteration, whereas testing of personalized medicine (Figure 3b) investigates whether the concept of treatment individualization improves outcomes in 'all comers'. The screening of patients for inclusion in clinical trials is the major limitation for the development of stratified medicine in MBC, because most of the current candidate drivers are detectable in less than 10% of the patients. If we consider the development of FGFR1 inhibitors for example, and assume the same statistical analyses as in the BOLERO-2 trial,<sup>83</sup> in which a total of 528 progression-free survival (PFS) events were required in order to generate a hazard ratio of 0.74 with 90% power, and assuming a median PFS of 3.7 months in the control group, and a 2:1 randomization ratio in favour of the investigational arm, a total 705 patients would need to be randomized. If we further assume that *FGFR1*-amplification is observed in around 10% patients and a screening failure rate of around 15%, then a total of 8,294 patients will be required to be screened to make this study feasible.

This approach generates two major issues: first, when a patient is tested for a single gene alteration, its likelihood of being positive and, therefore, treated by a drug matched to a genomic alteration is very low. One solution to address this issue is to use high-throughput genomic approaches to detect all present genomic alterations and enable the patient to be assigned to a specific therapeutic trial. The second issue is the need to screen a large number of patients to perform a clinical trial. If we take the scenario of drug development in an *AKT1*-mutant population (4% of all breast cancers), there is a need to screen around 10,000 patients with MBC to be able to perform a randomized clinical trial that includes 400 patients.

### Drug development in rare genomic segments

To overcome the issue of patient accrual in a stratified medicine trial, we propose five possible solutions (Figure 4). First, a need to scale-up the number of patients screened for a genomic alteration exists. One possible approach consists of developing large molecular screening programmes to feed downstream therapeutic trials with patients presenting with a candidate genomic alteration. Two different types of molecular screening programmes for stratified medicine are available: basket trials and umbrella trials. The basket trials test the effect of a single drug on a molecular alteration in a variety of cancers. This design allows not only for a faster identification of candidate patients, but also to assess the potential value of this targeted therapy across different tumour types. The most informative example for breast cancer is a clinical trial that examined the administration of neratinib in patients



**Figure 4** | Overcoming the accrual challenges of stratified medicine approaches. Different potential solutions include scaling-up molecular screening by using either less invasive and more accessible methods, such as ctDNA, or creating large international screening programmes; the development of novel preclinical models; the use of the preoperative setting in drug development to validate potential biomarkers or to determine the biological effect of a drug; and grant drug approval based on phase II clinical trials in rare genomics segments with worse outcome. Abbreviations: BC, breast cancer; ctDNA, circulating tumour DNA.

who have solid tumours with activating *ERBB* mutations, including *EGFR*, *HER2* and *HER3*. Several distinct cancer cohorts are included in this study: bladder/urinary tract cancer, breast cancer, colorectal cancer, endometrial cancer, gastric/esophageal cancer, ovarian cancer, all other solid tumours with a *HER2* mutation, *EGFR* mutated and/or amplified primary brain cancer, and solid tumours with a *HER3* mutation (NCT01953926).<sup>84</sup> This trial will define whether patients with *ERBB2*-mutated breast cancer are sensitive to the pan-HER inhibitor neratinib. Preliminary data for the non-small-cell lung cancer (NSCLC) cohort showed that for the 13 patients in the trial who received neratinib monotherapy, no patient experienced a partial response (PR), 7 (54%) patients achieved stable disease (SD) and 4 (31%) patients achieved clinical benefit (defined as a PR or SD for 12 or more weeks).<sup>85</sup> Other basket trials have been developed recently, but are less relevant for breast cancer, owing to the scarcity of the molecular alterations that are investigated. These trials include the evaluation of vemurafenib in *BRAF*<sup>V600E</sup>-positive cancers (NCT01524978)<sup>86</sup> or crizotinib in patients with molecular alterations in *ALK*, *MET* or *ROS* genes (NCT02034981).<sup>87</sup> The umbrella trials assess the effect of different drugs in different molecular alterations either in one or several tumours. Two umbrella studies have been specifically dedicated to breast cancer, the SAFIR01 and the AURORA trials. The SAFIR01 trial included 403 patients that assigned 52 (13%) patients to a targeted therapy matched to an identified genomic alteration.<sup>25</sup> The AURORA trial is an ongoing molecular screening programme that will include 1,300 patients with advanced-stage breast cancer to perform high coverage NGS and RNA sequencing on matched primary and metastatic samples.<sup>88</sup> Other examples of molecular screening programmes that include breast cancers among other disease subtypes are also available. The NCI-MATCH

programme will include 3,000 patients with different types of cancer in order to find early signals of a response to targeted therapies. Finally, the best illustration of a molecular screening programme for drug registration is the Lung-MAP MASTER protocol<sup>89</sup> for patients with lung cancer. This trial will screen 10,000 patients to identify those who are suitable for inclusion in one of five randomized registration trials. So far, an example of such a large trial does not exist in breast cancer as most of the molecular screening programmes are developed by single institutions for inclusion in their own phase I trials.<sup>90</sup> The only two large molecular screening programmes available are the previously mentioned SAFIR01 trial for breast cancer patients<sup>25</sup> and the BATTLE trial for NSCLC,<sup>91</sup> both of them in a single tumour type.

A second solution to circumvent low incidence of targetable alterations consists of clustering several genomic alterations as a single predictor. As an illustration, in a randomized trial testing the *FGFR*-inhibitor dovitinib, *FGFR1*, *FGFR2* and *FGF3* amplifications were clustered within a single population called “FGF-pathway amplified breast cancer.”<sup>34</sup> Following a similar principle, *BRCA1* and *BRCA2*-mutation carriers are usually clustered together in clinical trials testing PARP inhibitors.<sup>92</sup>

The third approach is to perform most of the drug development in the preoperative setting. The drug development process usually includes several trials for proof-of-concept, dose optimization, biomarker validation and development of combination therapies. Overall early stage breast cancer represents >80% of the breast cancers detected, thus several trials could be performed in the preoperative setting rather than in patients with metastatic disease, which would speed up the process.

A fourth potential solution would be the use of novel preclinical models (such as patient-derived xenografts and *ex-vivo* circulating tumour cell [CTC] cultures). As an example, studies using *ex vivo* culture of circulating breast tumour cells for individualized testing of drug susceptibility have been performed,<sup>93</sup> including a trial in CTC cultures from six patients with ER-positive breast cancer that were tumorigenic in mice, allowing for genome sequencing of the CTC lines but also drug sensitivity testing that revealed potential new therapeutic targets.<sup>93</sup> The use of CTCs would enable the use of genome sequencing to identify the presence of pre-existing or newly acquired molecular alterations in the tumour, and enables drug sensitivity testing to identify the best therapies for individual cancer patients.

Furthermore, drug approval based on phase I/II trials could dramatically speed-up drug development, as shown by the example of *ALK* inhibitors in lung cancer.<sup>88</sup> This approach would require the initial identification of molecular subgroups associated with very poor outcome. *ESR1* mutations have been associated with very poor outcome in patients with MBC<sup>14</sup> and could constitute a population where drugs could be approved based on the results of phase II clinical trials. Of note, this strategy is effective only if the genomic segment associated with poor outcome also defines a population that is sensitive to the investigated targeted therapy.

### Feasibility of biopsies and genomic tests

While accrual represents the biggest challenge to develop stratified medicine for patients with breast cancer, other logistical considerations need to be addressed. The most relevant one is that it is impossible to perform genomic tests in a significant proportion of patients, owing to difficult tumour (metastatic) locations (such as bone metastasis)<sup>94</sup> and the low percentage of cancer cells in some samples. Several approaches to overcome these limitations are available. Firstly, the analysis of ctDNA<sup>29</sup> could replace the use of biopsies when these cannot be done, or when the tumour site is not compatible with DNA extraction. However, there is lack of robustness to quantify copy-number aberrations using this technology. Performing a biopsy might be an issue not only in patients with tumours in difficult-to-access locations, and it seems likely that the capacity of interventional radiology facilities will not permit biopsy sampling in all patients. Secondly, performing in-depth sequencing could enable analyses of samples with a low percentage of cancer cells, as shown in the MOSCATO trial, where the introduction of NGS (1,000×) increased the feasibility of widespread genomic testing.<sup>15</sup>

### Clinical utility of multigene profiling in MBC

Stratified medicine is currently the dominant model in breast cancer research, although, its sustainability in the long term is questionable. Indeed, with the evolution of technologies and advances in biology, rarer genomic entities will be detected, leading to ever more challenges for patient accrual and will make stratified clinical trials unfeasible. A more-appropriate model to develop precision medicine in patients with MBC could consist of evaluating the concept of personalized medicine, rather than stratified medicine. In this model, the hypothesis of clinical research is that the use of a multigene profiling will improve outcomes in the whole population (Figure 3b). These trials compare a standard treatment arm to an arm in which stratification is tested based on molecular profiling. In patients with breast cancer, the SAFIR02 trial<sup>95</sup> is being conducted to compare maintenance chemotherapy to eight different targeted therapies given according to an individual's genomic profile (identified by NGS and CGH arrays). This trial, therefore, evaluates whether the overall approach of personalized medicine improves outcome as opposed to a standard maintenance therapy. Several other studies are testing this strategy; most of them including multiple tumour types.<sup>90</sup> The main advantage of this approach is that it does not need to accrue thousands of patients in order to address a research question. Nevertheless, several key concerns about trial design remain unsolved. The first one relates to the choice of the standard arm. Indeed, the best standard arm should theoretically include similar treatments to those used in the personalized medicine arm, but given in a random way, in order to have the same drugs both in the standard and the personalized-treatment arm. This strategy is adopted in the NCI-MPACT trial<sup>96</sup> testing efficacy of personalized medicine across tumour types, including breast cancers. The

second question relates to whether the effect observed in the overall population is an effect that is common all the drugs tested or only a few drugs. Thus, several statistical tools need to be developed to exclude this bias, such as interaction tests to exclude the hypothesis that some treatments are better than others. The sample sizes of these trials could be calculated accordingly.

Other logistical and operational challenges exist, but they are currently not considered as major limiting factors to developing precision medicine, including the fact that a genomic alteration cannot be identified in every patient. In this context, the use of a large panel of genes could maximize the likelihood of a patient receiving a matched therapy. Also, developing alternative technologies to identify drivers, including gene expression or protein-based assays, would certainly increase the number of patients for which a driver is identified. For example, the WINTHER trial is performing gene-expression arrays in patients for which no DNA alteration could be detected to compare it to gene expression in normal (not cancerous) tissue to identify a potential treatment of interest.<sup>97</sup> The difficulty of access to appropriate or available drugs is an important secondary issue. Indeed, in the SAFIR01 study, only 28% of the patients with a genomic alteration had access to the targeted therapy,<sup>25</sup> either because a clinical trial testing a specific targeted agent would not accept patients with breast cancer or owing to the absence of available treatment slots. These issues indicate that genomic tests should be performed only in the context of a drug access programme, either in phase I units or in consortium with secured drug access. Finally, the modality of implementation of precision medicine should also be considered. Indeed, two models of precision medicine are available. In one model, some biomarker companies would perform the test whereas in the other one, public centres will conduct them for free. As an illustration of the latter, the French National Cancer Institute has set-up a group of 28 genomic centres located in academic hospitals to offer free and equal access to genomics for all patients with cancer.<sup>83</sup>

### Conclusions

The concept of precision medicine driven by genomics for MBC is appealing; however, no evidence exists that it can improve patient outcome. This failure is a result of both scientific and logistical issues. The biggest issue in the field of breast cancer is the lack of well characterized drivers, with the exceptions of *ER*, *ERBB2*, *PIK3CA* and *AKT1*. Developing catalogues of cancer-related genes, together with methods for assessments of pathway activation could enable a better identification of such drivers in the near future. The characterization of the genomic landscape of tumours and of the activated protein network will guide combination therapies to optimize therapeutic effects. Finally, logistics and operational challenges need to be addressed. In this regard, a shift towards personalized medicine rather than a stratified medicine approach could improve the management of patients.



1. Jensen, E. V. & DeSombre, E. R. Oestrogen-receptor interaction. *Science* **182**, 126–134 (1973).
2. Slamon, D. J. *et al.* Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* **235**, 177–182 (1987).
3. Osborne, C. K. Tamoxifen in the treatment of breast cancer. *N. Engl. J. Med.* **339**, 1609–1618 (1998).
4. Dawood, S., Broglio, K., Buzdar, A. U., Hortobagyi, G. N. & Giordano, S. H. Prognosis of women with metastatic breast cancer by HER2 status and trastuzumab treatment: an institutional-based review. *J. Clin. Oncol.* **28**, 92–98 (2010).
5. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333–339 (2013).
6. Stephens, P. J. *et al.* The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**, 400–404 (2012).
7. Hortobagyi, G. N. *et al.* Correlation of molecular alterations with efficacy of everolimus in hormone receptor–positive, HER2-negative advanced breast cancer: results from BOLERO-2 [abstract]. *J. Clin. Oncol.* **31** (Suppl.), LBA509 (2013).
8. Treilleux, I. *et al.* Translational studies within the TAMRAD randomized GINECO trial: evidence for mTORC1 activation marker as a predictive factor for everolimus efficacy in advanced breast cancer. *Ann. Oncol.* **26**, 120–125 (2015).
9. Shah, S. P. *et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486**, 395–399 (2012).
10. Amir, E. *et al.* Prospective study evaluating the impact of tissue confirmation of metastatic disease in patients with breast cancer. *J. Clin. Oncol.* **30**, 587–592 (2012).
11. Niikura, N. *et al.* Loss of human epidermal growth factor receptor 2 (HER2) expression in metastatic sites of HER2-overexpressing primary breast tumours. *J. Clin. Oncol.* **30**, 593–599 (2012).
12. Toy, W. *et al.* ESR1 ligand-binding domain mutations in hormone-resistant breast cancer. *Nat. Genet.* **45**, 1439–1445 (2013).
13. Robinson, D. R. *et al.* Activating ESR1 mutations in hormone-resistant metastatic breast cancer. *Nat. Genet.* **45**, 1446–1451 (2013).
14. Arnedos, M. *et al.* Genomic and immune characterization of metastatic breast cancer (mBC): And ancillary study of the SAFIRO1 & MOSCATO trials [abstract]. *Ann. Oncol.* **25**, a3510 (2014).
15. Massard, C. *et al.* Enriching phase I trials with molecular alterations: Interim analysis of 708 patients enrolled in the MOSCATO 01 trial [abstract]. *13th International Congress on Targeted Anticancer Therapies, Paris, France*, a03.7 (2015).
16. Lord, C. J. & Ashworth, A. Mechanisms of resistance to therapies targeting BRCA-mutant cancers. *Nat. Med.* **19**, 1381–1388 (2013).
17. Wagle, N. *et al.* Whole exome sequencing (WES) of HER2+ metastatic breast cancer (MBC) from patients with or without prior trastuzumab (T): A correlative analysis of TBCRC003. *Cancer Res.* **75**, PD3-PD5 (2015).
18. Hammond, M. E. *et al.* American Society of Clinical Oncology/College Of American Pathologists guideline recommendations for immunohistochemical testing of oestrogen and progesterone receptors in breast cancer. *J. Clin. Oncol.* **28**, 2784–2795 (2010).
19. Wolff, A. C. *et al.* Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update. *J. Clin. Oncol.* **31**, 3997–4013 (2013).
20. Paik, S. *et al.* A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.* **351**, 2817–2826 (2004).
21. van de Vijver, M. J. *et al.* A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **347**, 1999–2009 (2002).
22. Sorlie, T. *et al.* Repeated observation of breast tumour subtypes in independent gene expression data sets. *Proc. Natl Acad. Sci. USA* **100**, 8418–8423 (2003).
23. Loi, S. *et al.* PIK3CA genotype and a PIK3CA mutation-related gene signature and response to everolimus and letrozole in oestrogen receptor positive breast cancer. *PLoS ONE* **8**, e53292 (2013).
24. Balko, J. M. *et al.* A gene expression signature of MEK pathway activation to predict survival in triple-negative breast cancer [abstract]. *J. Clin. Oncol.* **30** (Suppl.), a1024 (2012).
25. Andre, F. *et al.* Comparative genomic hybridisation array and DNA sequencing to direct treatment of metastatic breast cancer: a multicentre, prospective trial (SAFIRO1/UNICANCER). *Lancet Oncol.* **15**, 267–274 (2014).
26. Arnedos, M. *et al.* Array CGH and PIK3CA/AKT1 mutations to drive patients to specific targeted agents: a clinical experience in 108 patients with metastatic breast cancer. *Eur. J. Cancer* **48**, 2293–2299 (2012).
27. Van Allen, E. M. *et al.* Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumour samples to guide precision cancer medicine. *Nat. Med.* **20**, 682–688 (2014).
28. Watkins, J. A., Irshad, S., Grigoriadis, A. & Tutt, A. N. Genomic scars as biomarkers of homologous recombination deficiency and drug response in breast and ovarian cancers. *Breast Cancer Res.* **16**, 211 (2014).
29. Dawson, S. J. *et al.* Analysis of circulating tumour DNA to monitor metastatic breast cancer. *N. Engl. J. Med.* **368**, 1199–1209 (2013).
30. Murtaza, M. *et al.* Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. *Nature* **497**, 108–112 (2013).
31. Garnett, M. J. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570–575 (2012).
32. Cancer Genome Atlas, N. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
33. Janku, F. *et al.* PI3K/AKT/mTOR inhibitors in patients with breast and gynecologic malignancies harbouring PIK3CA mutations. *J. Clin. Oncol.* **30**, 777–782 (2012).
34. Andre, F. *et al.* Targeting FGFR with dovitinib (TKI258): preclinical and clinical data in breast cancer. *Clin. Cancer Res.* **19**, 3693–3702 (2013).
35. Soria, J. C. *et al.* Phase I/IIa study evaluating the safety, efficacy, pharmacokinetics, and pharmacodynamics of lucitanib in advanced solid tumours. *Ann. Oncol.* **25**, 2244–2251 (2014).
36. Bose, R. *et al.* Activating HER2 mutations in HER2 gene amplification negative breast cancer. *Cancer Discov.* **3**, 224–237 (2013).
37. Finn, R. S. *et al.* The cyclin-dependent kinase 4/6 inhibitor palbociclib in combination with letrozole versus letrozole alone as first-line treatment of oestrogen receptor-positive, HER2-negative, advanced breast cancer (PALOMA-1/TRIO-18): a randomised phase 2 study. *Lancet Oncol.* **16**, 25–35 (2015).
38. Malumbres, M. & Barbacid, M. Cell cycle, CDKs and cancer: a changing paradigm. *Nat. Rev. Cancer* **9**, 153–166 (2009).
39. Zoncu, R., Efeyan, A. & Sabatini, D. M. mTOR: from growth signal integration to cancer, diabetes and ageing. *Nat. Rev. Mol. Cell. Biol.* **12**, 21–35 (2011).
40. Dickler, M. *et al.* A first-in-human phase I study to evaluate the oral selective oestrogen receptor degrader GDC-0810 (ARN-810) in postmenopausal women with oestrogen receptor+ HER2–, advanced/metastatic breast cancer. *AACR [abstract CT231]* (2015).
41. Juric, D. *et al.* Convergent loss of PTEN leads to clinical resistance to a PI(3)Kalpha inhibitor. *Nature* **518**, 240–244 (2015).
42. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
43. Henderson, S., Chakravarthy, A., Su, X., Boshoff, C. & Fenton, T. R. APOBEC-mediated cytosine deamination links PIK3CA helical domain mutations to human papillomavirus-driven tumour development. *Cell Rep.* **7**, 1833–1841 (2014).
44. Swisher, E. *et al.* ARIEL2: A phase 2 study to prospectively identify ovarian cancer patients likely to respond to rucaparib [abstract]. *26th EORTC-NCI-AACR Symposium on Molecular Targets and Cancer Therapeutics*, a215 (2014).
45. Tutt, A. *et al.* Oral poly(ADP-ribose) polymerase inhibitor olaparib in patients with BRCA1 or BRCA2 mutations and advanced breast cancer: a proof-of-concept trial. *Lancet* **376**, 235–244 (2010).
46. Cui, Y., Palli, S. S., Innes, C. L. & Paules, R. S. Depletion of ATR selectively sensitizes ATM-deficient human mammary epithelial cells to ionizing radiation and DNA-damaging agents. *Cell Cycle* **13**, 541–550 (2014).
47. Weber, A. M. & Ryan, A. J. ATM and ATR as therapeutic targets in cancer. *Pharmacol. Ther.* **149**, 124–138 (2015).
48. Bartlett, J. M. *et al.* Predicting anthracycline benefit: TOP2A and CEP17-not only but also. *J. Clin. Oncol.* **33**, 1680–1687 (2015).
49. Nanda, R. *et al.* A phase Ib study of pembrolizumab (MK-3475) in patients with advanced triple-negative breast cancer [abstract]. *Cancer Res.* **75**, S1-09 (2015).
50. Snyder, A. *et al.* Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N. Engl. J. Med.* **371**, 2189–2199 (2014).
51. Seliger, B. *et al.* Immune escape of melanoma: first evidence of structural alterations in two distinct components of the MHC class I antigen processing pathway. *Cancer Res.* **61**, 8647–8650 (2001).
52. Tume, P. C. *et al.* PD-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature* **515**, 568–571 (2014).
53. Ignatiadis, M. *et al.* Gene modules and response to neoadjuvant chemotherapy in breast cancer subtypes: a pooled analysis. *J. Clin. Oncol.* **30**, 1996–2004 (2012).
54. Vacchelli, E. *et al.* Loss-of-function alleles of P2RX7 and TLR4 fail to affect the response to chemotherapy in non-small cell lung cancer. *Onco-immunology* **1**, 271–278 (2012).
55. Bendell, J. C. *et al.* Phase I, dose-escalation study of BKM120, an oral pan-class I PI3K inhibitor, in patients with advanced solid tumours. *J. Clin. Oncol.* **30**, 282–290 (2012).

56. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
57. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
58. Taylor, B. S. *et al.* Functional copy-number alterations in cancer. *PLoS ONE* **3**, e3179 (2008).
59. Dees, N. D. *et al.* MuSiC: identifying mutational significance in cancer genomes. *Genome Res.* **22**, 1589–1598 (2012).
60. Hua, X. *et al.* DrGaP: a powerful tool for identifying driver genes and pathways in cancer sequencing studies. *Am. J. Hum. Genet.* **93**, 439–451 (2013).
61. Gonzalez-Perez, A. *et al.* IntOGen-mutations identifies cancer drivers across tumour types. *Nat. Methods* **10**, 1081–1082 (2013).
62. Rubio-Perez, C. *et al.* *In silico* prescription of anticancer drugs to cohorts of 28 tumour types reveals targeting opportunities. *Cancer Cell* **27**, 382–396 (2015).
63. Suo, C. *et al.* Integration of somatic mutation, expression and functional data reveals potential driver genes predictive of breast cancer survival. *Bioinformatics* <https://dx.doi.org/10.1093/bioinformatics/btv164> (2015).
64. Andre, F. *et al.* Everolimus for women with trastuzumab-resistant, HER2-positive, advanced breast cancer (BOLERO-3): a randomised, double-blind, placebo-controlled phase 3 trial. *Lancet Oncol.* **15**, 580–591 (2014).
65. Andre, F. *et al.* Expression patterns and predictive value of phosphorylated AKT in early-stage breast cancer. *Ann. Oncol.* **19**, 315–320 (2008).
66. Mayer, I. A. *et al.* Stand up to cancer phase Ib study of pan-phosphoinositide-3-kinase inhibitor buparlisib with letrozole in oestrogen receptor-positive/human epidermal growth factor receptor 2-negative metastatic breast cancer. *J. Clin. Oncol.* **32**, 1202–1209 (2014).
67. Fritsch, C. *et al.* Characterization of the novel and specific PI3K $\alpha$  inhibitor NVP-BYL719 and development of the patient stratification strategy for clinical trials. *Mol. Cancer Ther.* **13**, 1117–1129 (2014).
68. Janku, F. *et al.* Phase I study of the PI3K $\alpha$  inhibitor BYL719 plus fulvestrant in patients with PIK3CA-altered and wild type ER+/HER2– locally advanced or metastatic breast cancer [abstract]. *Cancer Res.* **74**, PD5–5 (2014).
69. Fu, Y., Ferte, C., Soria, J. C., F., A. & Lefebvre, C. Combination of targetable gene alterations decreases anti-cancer drug treatment response in cell lines and patients [abstract P10]. *Cancer Pharmacogenomics and Targeted Therapies*, Hinxton, Cambridge, UK (2014).
70. Fontanella, C. *et al.* Does toxicity predict efficacy? Insight into the mechanism of action of lapatinib. *J. Clin. Oncol.* **32**, 3458–3459 (2014).
71. Baselga, J. *et al.* Biomarker analyses in CLEOPATRA: a phase III, placebo-controlled study of pertuzumab in human epidermal growth factor receptor 2-positive, first-line metastatic breast cancer. *J. Clin. Oncol.* **32**, 753–761 (2014).
72. Schneeweiss, A. *et al.* Evaluating the predictive value of biomarkers for efficacy outcomes in response to pertuzumab- and trastuzumab-based therapy: an exploratory analysis of the TRYPHAENA study. *Breast Cancer Res.* **16**, R73 (2014).
73. Lito, P., Rosen, N. & Solit, D. B. Tumour adaptation and resistance to RAF inhibitors. *Nat. Med.* **19**, 1401–1409 (2013).
74. Choi, Y. L. *et al.* EML4-ALK mutations in lung cancer that confer resistance to ALK inhibitors. *N. Engl. J. Med.* **363**, 1734–1739 (2010).
75. Eirew, P. *et al.* Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature* **518**, 422–426 (2015).
76. Wang, Y. *et al.* Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* **512**, 155–160 (2014).
77. Kosaka, K. *et al.* Emergence of resistant variants detected by ultra-deep sequencing after asunaprevir and daclatasvir combination therapy in patients infected with hepatitis C virus genotype 1. *J. Viral. Hepat.* **22**, 158–165 (2015).
78. De Mattos-Arruda, L. *et al.* Capturing intra-tumour genetic heterogeneity by *de novo* mutation profiling of circulating cell-free tumour DNA: a proof-of-principle. *Ann. Oncol.* **25**, 1729–1735 (2014).
79. O'Reilly, K. E. *et al.* mTOR inhibition induces upstream receptor tyrosine kinase signalling and activates Akt. *Cancer Res.* **66**, 1500–1508 (2006).
80. Atzori, F. *et al.* A phase I pharmacokinetic and pharmacodynamic study of dalotuzumab (MK-0646), an anti-insulin-like growth factor-1 receptor monoclonal antibody, in patients with advanced solid tumours. *Clin. Cancer Res.* **17**, 6304–6312 (2011).
81. Chen, X. *et al.* Dual inhibition of PI3K and mTOR mitigates compensatory AKT activation and improves tamoxifen response in breast cancer. *Mol. Cancer Res.* **11**, 1269–1278 (2013).
82. US National Library of Medicine. *ClinicalTrials.gov* [online], <https://www.clinicaltrials.gov/ct2/show/NCT02129556> (2015).
83. Baselga, J. *et al.* Everolimus in postmenopausal hormone-receptor-positive advanced breast cancer. *N. Engl. J. Med.* **366**, 520–529 (2012).
84. US National Library of Medicine. *ClinicalTrials.gov* [online], <https://www.clinicaltrials.gov/ct2/show/NCT01953926> (2015).
85. Besse, B. *et al.* Neratinib (N) with or without temsirolimus (TEM) in patients (pts) with non-small cell lung cancer (NSCLC) carrying HER2 somatic mutations: an international randomized phase II study [abstract]. *Ann. Oncol.* **25**, LBA39 (2014).
86. Hyman, D. M. *et al.* VE-BASKET, a first-in-kind, phase II, histology-independent “basket” study of vemurafenib (VEM) in nonmelanoma solid tumours harbouring BRAF V600 mutations (V600m) [abstract]. *J. Clin. Oncol.* **32** (Suppl. 5), a2533 (2014).
87. US National Library of Medicine. *ClinicalTrials.gov* [online], <https://www.clinicaltrials.gov/ct2/show/NCT02034981> (2015).
88. Zardavas, D. *et al.* The AURORA initiative for metastatic breast cancer. *Br. J. Cancer* **111**, 1881–1887 (2014).
89. US National Library of Medicine. *ClinicalTrials.gov* [online], <https://www.clinicaltrials.gov/ct2/show/NCT02154490> (2015).
90. Bedard, P. L., Hansen, A. R., Ratain, M. J. & Siu, L. L. Tumour heterogeneity in the clinic. *Nature* **501**, 355–364 (2013).
91. Kim, E. S. *et al.* The BATTLE trial: personalizing therapy for lung cancer. *Cancer Discov.* **1**, 44–53 (2011).
92. Kaufman, B. *et al.* Olaparib monotherapy in patients with advanced cancer and a germline BRCA1/2 mutation. *J. Clin. Oncol.* **33**, 244–250 (2015).
93. Yu, M. *et al.* Cancer therapy. *Ex vivo* culture of circulating breast tumour cells for individualized testing of drug susceptibility. *Science* **345**, 216–220 (2014).
94. Singh, V. M. *et al.* Analysis of the effect of various decalcification agents on the quantity and quality of nucleic acid (DNA and RNA) recovered from bone biopsies. *Ann. Diagn. Pathol.* **17**, 322–326 (2013).
95. US National Library of Medicine. *ClinicalTrials.gov* [online], <https://www.clinicaltrials.gov/ct2/show/NCT02299999> (2014).
96. US National Library of Medicine. *ClinicalTrials.gov* [online], <https://www.clinicaltrials.gov/ct2/show/NCT01827384> (2015).
97. US National Library of Medicine. *ClinicalTrials.gov* [online], <https://www.clinicaltrials.gov/ct2/show/NCT01856296> (2013).

## Acknowledgements

S.L. is supported by Cancer Council Victoria and the National Health and Medical Research Council of Australia (NHMRC). F.A. is supported by Grants from the ARC, Breast cancer Research Foundation, Odyssey and Operation Parrains Chercheurs.

## Author contributions

H.B. and F.A. developed the outline; F.A. wrote the sections related to genomics and modalities of development. All authors made a significant contribution to writing the manuscript, reviewing and/or editing the manuscript and approved the final version before submission.

ARTICLE

Received 29 Aug 2014 | Accepted 28 Nov 2014 | Published 9 Jan 2015

DOI: 10.1038/ncomms6987

OPEN

# *BCL11A* is a triple-negative breast cancer gene with critical functions in stem and progenitor cells

Walid T. Khaled<sup>1,2,\*</sup>, Song Choon Lee<sup>1,\*</sup>, John Stingl<sup>3</sup>, Xiongfeng Chen<sup>4</sup>, H. Raza Ali<sup>3,5</sup>, Oscar M. Rueda<sup>3</sup>, Fazal Hadi<sup>2</sup>, Juexuan Wang<sup>1</sup>, Yong Yu<sup>1</sup>, Suet-Feung Chin<sup>3</sup>, Mike Stratton<sup>1</sup>, Andy Futreal<sup>1</sup>, Nancy A. Jenkins<sup>6</sup>, Sam Aparicio<sup>7</sup>, Neal G. Copeland<sup>6</sup>, Christine J. Watson<sup>8</sup>, Carlos Caldas<sup>3,5,9</sup> & Pentao Liu<sup>1</sup>

Triple-negative breast cancer (TNBC) has poor prognostic outcome compared with other types of breast cancer. The molecular and cellular mechanisms underlying TNBC pathology are not fully understood. Here, we report that the transcription factor *BCL11A* is overexpressed in TNBC including basal-like breast cancer (BLBC) and that its genomic locus is amplified in up to 38% of BLBC tumours. Exogenous *BCL11A* overexpression promotes tumour formation, whereas its knockdown in TNBC cell lines suppresses their tumourigenic potential in xenograft models. In the DMBA-induced tumour model, *Bcl11a* deletion substantially decreases tumour formation, even in p53-null cells and inactivation of *Bcl11a* in established tumours causes their regression. At the cellular level, *Bcl11a* deletion causes a reduction in the number of mammary epithelial stem and progenitor cells. Thus, *BCL11A* has an important role in TNBC and normal mammary epithelial cells. This study highlights the importance of further investigation of *BCL11A* in TNBC-targeted therapies.

<sup>1</sup> Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1HH, UK. <sup>2</sup> Department of Pharmacology, University of Cambridge, Cambridge CB2 1PD, UK. <sup>3</sup> Cancer Research UK Cambridge Institute, and Department of Oncology, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK. <sup>4</sup> SAIC-Frederic, National Cancer Institute-Frederick, Frederick, Maryland 21701, USA. <sup>5</sup> Cambridge Experimental Cancer Medicine Centre, Cambridge CB2 0RE, UK. <sup>6</sup> The Methodist Hospital Research Institute, 6670 Bertner Street, Houston, Texas 77030, USA. <sup>7</sup> Molecular Oncology Department, BC Cancer Agency Research Centre, 675 West 10th Avenue, Vancouver, British Columbia V5Z 1L3, Canada. <sup>8</sup> Department of Pathology, University of Cambridge, Cambridge CB2 1QP, UK. <sup>9</sup> Addenbrooke's Hospital, Cambridge University Hospital NHS Foundation Trust and NIHR Cambridge Biomedical Research Centre, Cambridge CB2 2QQ, UK. \*These authors contributed equally to this work. Correspondence and requests for materials should be addressed to W.T.K. (email: wtk22@cam.ac.uk) or to P.L. (email: pl2@sanger.ac.uk).



One of the major challenges in treating breast cancer is the heterogeneous nature of the disease<sup>1</sup>. TNBC accounts for around 15% of all breast cancer cases and in the absence of effective targeted therapies, TNBC patients tend to have a poor prognosis<sup>2–4</sup>. At the molecular level, several distinct subtypes of breast cancer have been identified based on the gene expression profiling<sup>3,5,6</sup>. The most commonly used classification describes six subtypes: luminal A, luminal B, Her2, claudin low, basal-like breast cancer (BLBC) and normal<sup>3,6</sup>. More recently, analysis of large numbers of tumour samples as part of the METABRIC study identified 10 pathologically distinct subtypes known as integrative cluster (IC) 1–10 (ref. 5). The majority of TNBC cases (80%) have a BLBC<sup>7</sup> or IC10 (ref. 5) gene expression signatures. In addition, cancer sequencing studies have identified mutations of *p53*, *PTEN* and *BRCA1* in TNBC<sup>2,4,8,9</sup>. However, driver oncogenic genomic aberrations in TNBC have not been comprehensively identified.

The developmental hierarchies of the mammary epithelium and hematopoietic lineages share many similarities<sup>10</sup> in that stem cells progressively give rise to lineage-restricted progenitors, which ultimately differentiate and generate all functional cells. A number of key hematopoiesis transcription factors have important roles in mouse mammary gland development and are human breast cancer genes<sup>11–15</sup>. For example, the key regulator of T-helper-2 cell development, *GATA3*, is critical in luminal mammary cell development<sup>12,13</sup> and is a luminal breast cancer marker gene<sup>16</sup>. In this study we interrogated cancer genomics data focusing on a subset of important hematopoiesis factors and identified *BCL11A* as a novel TNBC oncogene.

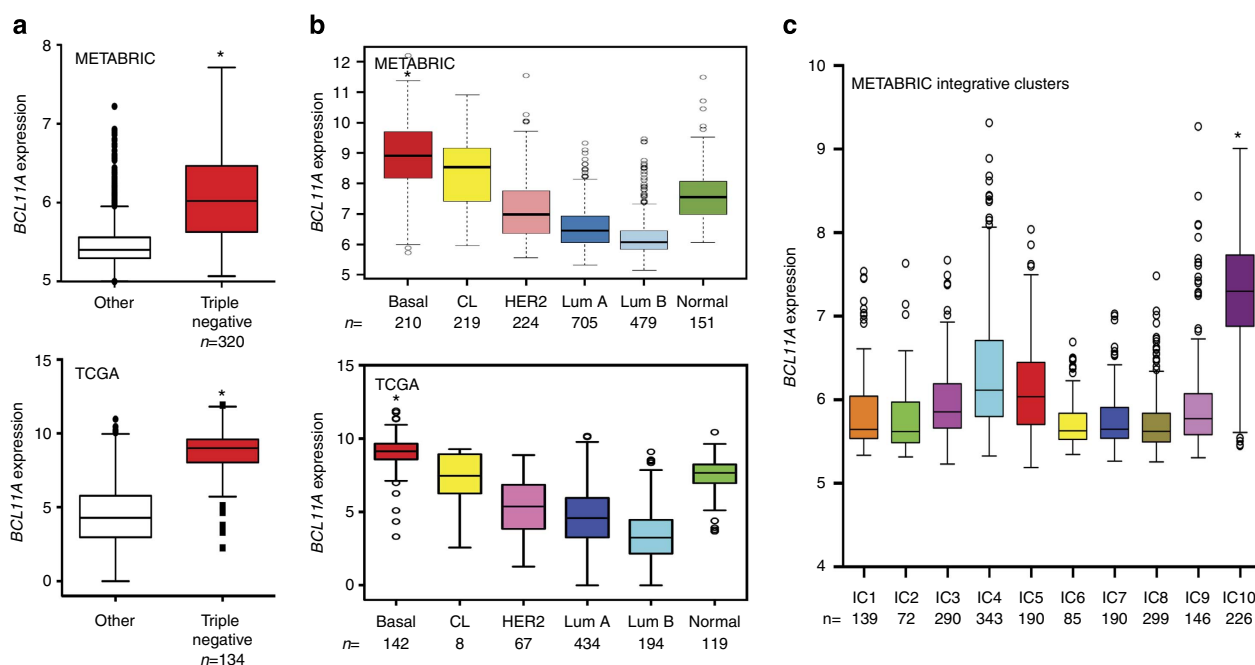
## Results

***BCL11A* is highly expressed in triple-negative breast cancer.** In an attempt to identify potential TNBC oncogenes, we selected a list of genes known to have important roles in hematopoiesis and investigated their expression across the major molecular

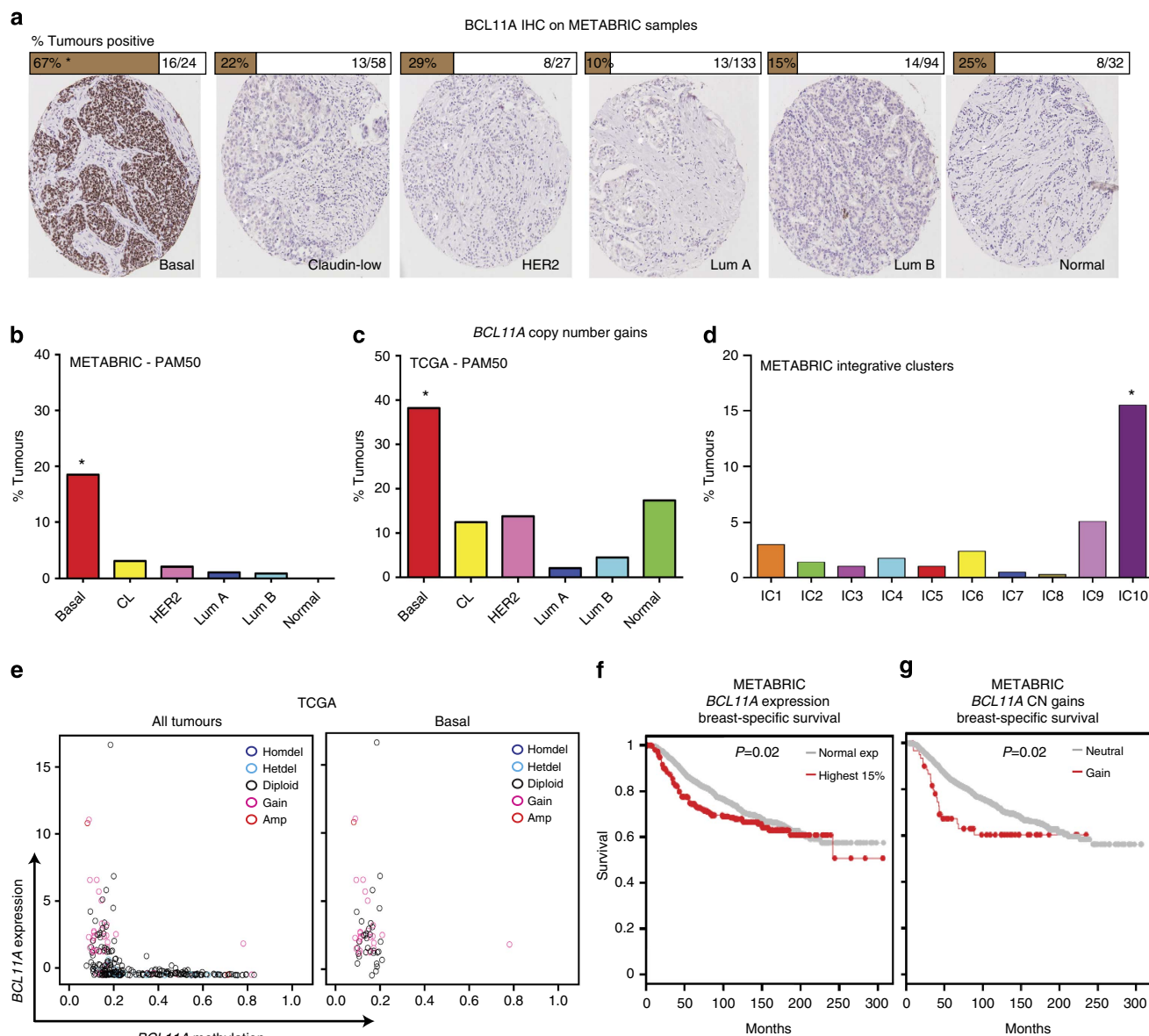
subtypes of breast cancer<sup>3</sup>. We first re-analysed a publically available microarray data set<sup>6</sup> and found that out of the examined genes, *BCL11A* was differentially and highly expressed in BLBC (Supplementary Fig. 1a). This is in sharp contrast to *GATA3*, which is highly expressed only in luminal subtypes (Supplementary Fig. 1a) and is a known prognostic marker for these tumours<sup>16</sup>.

We then investigated the expression of *BCL11A* in other patient data sets including METABRIC<sup>5</sup> and TCGA<sup>8</sup>, which between them have curated gene expression, copy number (CN) variation and clinical data from close to 3,000 patients<sup>5</sup>. Pathologically, we found that high *BCL11A* expression significantly correlated with TNBC pathology (Fig. 1a). At the molecular level, high *BCL11A* expression was also found to significantly correlate with the BLBC subtype in the METABRIC, TCGA and six other microarray data sets (Fig. 1b and Supplementary Fig. 1b). Quantitative reverse transcription PCR (qRT-PCR) analysis of *BCL11A* expression on a randomly selected subset of METABRIC tumours (all subtypes,  $n = 230$ ) validated the above expression data (Supplementary Fig. 2a). In addition, we also found that high *BCL11A* expression in METABRIC samples correlated with the recently described IC10 cluster of tumours (Fig. 1c), thus further supporting the concordance between the BLBC and IC10 classifications. Consistent with TNBC cases, high *BCL11A* expression was significantly correlated with a high histological grade (Supplementary Fig. 2b).

Furthermore, high *BCL11A* expression in BLBC cases was further validated by immunohistochemistry (IHC) on a subset of the METABRIC tumours (all subtypes,  $n = 368$ . BLBC,  $n = 24$ ). Strong *BCL11A* immunostaining was predominantly found in BLBC (Fig. 2a). Out of 24 BLBC samples examined from this subset, 16 scored positive for *BCL11A* (Fig. 2a; details in Methods). In addition, samples stained positively in IHC also had higher RNA levels compared with those scored as negative (Supplementary Fig. 2c).



**Figure 1 | *BCL11A* is highly expressed in TNBC.** (a) Significant correlation between *BCL11A* expression and the TNBC type of breast cancer in both METABRIC ( $n = 2,000$ ) and TCGA ( $n = 1,100$ ) data sets—\* indicates  $t$ -test  $P$  value  $< 0.005$ . (b) *BCL11A* expression across the six molecular subtypes of breast cancer ('Normal' refers to the PAM50 subtype) in both METABRIC and TCGA data sets—\* indicates  $t$ -test  $P$  value  $< 0.005$ . (c) The METABRIC samples distributed according to the ICs 1–10, showing the correlation between *BCL11A* expression and IC10—\* indicates  $t$ -test  $P$  value  $< 0.005$ .



**Figure 2 | Genomic alterations at the *BCL11A* locus.** (a) Images and scoring of BCL11A IHC staining on a subset of tumours from the METABRIC study (see Methods for scoring)—\* indicated  $\chi^2$ -test  $P < 0.0001$ . (b–d) Bar chart depicting the percentage of samples that harbour *BCL11A* CN gains in each subtype in both METABRIC and TCGA data sets—\* indicates  $\chi^2$ -test  $P < 0.0001$ . (e) Scatter plots showing the methylation status of the *BCL11A* locus in all tumours or basal only from the TCGA data set. The colour-coded legend indicated the *BCL11A* CN status for each tumour. (f–g) Kaplan-Meier plots showing the survival rate comparison between patients who have normal or high levels of *BCL11A* expression, or between patients with CN gains or without (neutral).

One mechanism for the induction of high *BCL11A* expression in BLBC cases could be CN aberrations. From ~2,000 breast cancer cases in METABRIC<sup>5</sup>, CN gains at the *BCL11A* genomic locus were identified in 62 patients (Supplementary Fig. 3a), which also correlates with high *BCL11A* expression (Supplementary Fig. 3b). Importantly, out of these 62 patients with CN gains, 39 were classified as BLBC, which account for 18.6% (39/210) of the total BLBC cases in METABRIC (Fig. 2b). Examination of the TCGA data set revealed that 38% (31/81) of BLBC samples have *BCL11A* CN gains, which is again significantly correlated with higher gene expression (Fig. 2c and Supplementary Fig. 3c). A similar result was also found when the METABRIC data was analysed using the integrative clustering, with 15.6% of IC10 samples having *BCL11A* CN gains (Fig. 2d).

Further analysis of the TCGA data set revealed that in BLBCs, the *BCL11A* locus is almost exclusively hypomethylated and this is correlated with high expression levels (Fig. 2e). There was also no correlation between *BCL11A* CNs and the methylation status. This result suggests that epigenetic changes at the *BCL11A* locus could be another mechanism that contributes to its high expression in BLBC. Given the strong correlation with TNBC, patients with either high expression or CN gains of *BCL11A* had poor survival rates compared with the rest of the cohort (Fig. 2f, g). A similar trend was also observed in four other patient data sets<sup>17–20</sup> (Supplementary Fig. 3 f–i). In particular, patients with CN gains of *BCL11A* had a higher rate of relapse and metastasis and a lower rate of survival (Supplementary Fig. 3d–e). The utility of *BCL11A* expression/CN as a biomarker in the clinic thus warrants further investigation. Indeed, the future release of

patient outcome for the complete TCGA cohort will aid in clarifying this finding.

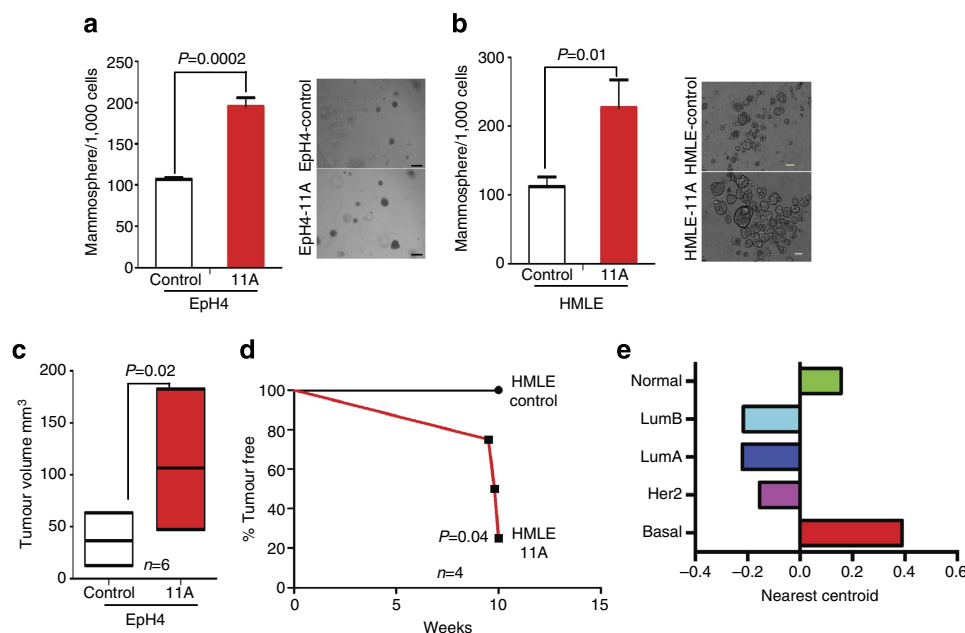
### High levels of *BCL11A* promote tumour development.

Although *BCL11A* is involved in rare B-cell lymphomas and is able to transform fibroblast cells *in vitro*<sup>21,22</sup>, the cellular and molecular mechanisms of *BCL11A*-mediated tumourigenesis remains unclear. To address this, we first tested whether *BCL11A* overexpression could promote the colony formation or tumour development in mammary epithelial cells. We overexpressed *BCL11A* in immortalized non-tumourigenic mouse Eph4 (ref. 23) or human HMLE<sup>24,25</sup> cells (Supplementary Fig. 4a) and performed Matrigel and suspension mammosphere assays. Forced *BCL11A* expression in both Eph4 and HMLE (Eph4-11A and HMLE-11A) cells resulted in double the number of spheres compared with their respective control cells (Fig. 3a–b). Furthermore, mouse Eph4-11A cells injected orthotopically in cleared mammary fat pads of immune-compromized *NOD/SCID/IL2 $\gamma$ <sup>-/-</sup>* (NSG) mice<sup>26</sup> formed larger and palpable tumours compared with control cells ( $n=6$ ) (Fig. 3c and Supplementary Fig. 4b). Similarly, three out of four mice injected with HMLE-11A cells developed tumours within 8 weeks of injection (Fig. 3d and Supplementary Fig. 4c) suggesting that elevated levels of *BCL11A* promote tumour development. Moreover, gene expression analysis of these three tumours along with the 2,000 tumours from the METABRIC study clustered them with the BLBC subgroup (Fig. 3e).

**Knockdown of *BCL11A* reduces tumourigenicity of TNBC cells.** Analysis of *BCL11A* expression in a panel of breast cancer cell lines revealed that *BCL11A* is highly expressed in TNBC lines

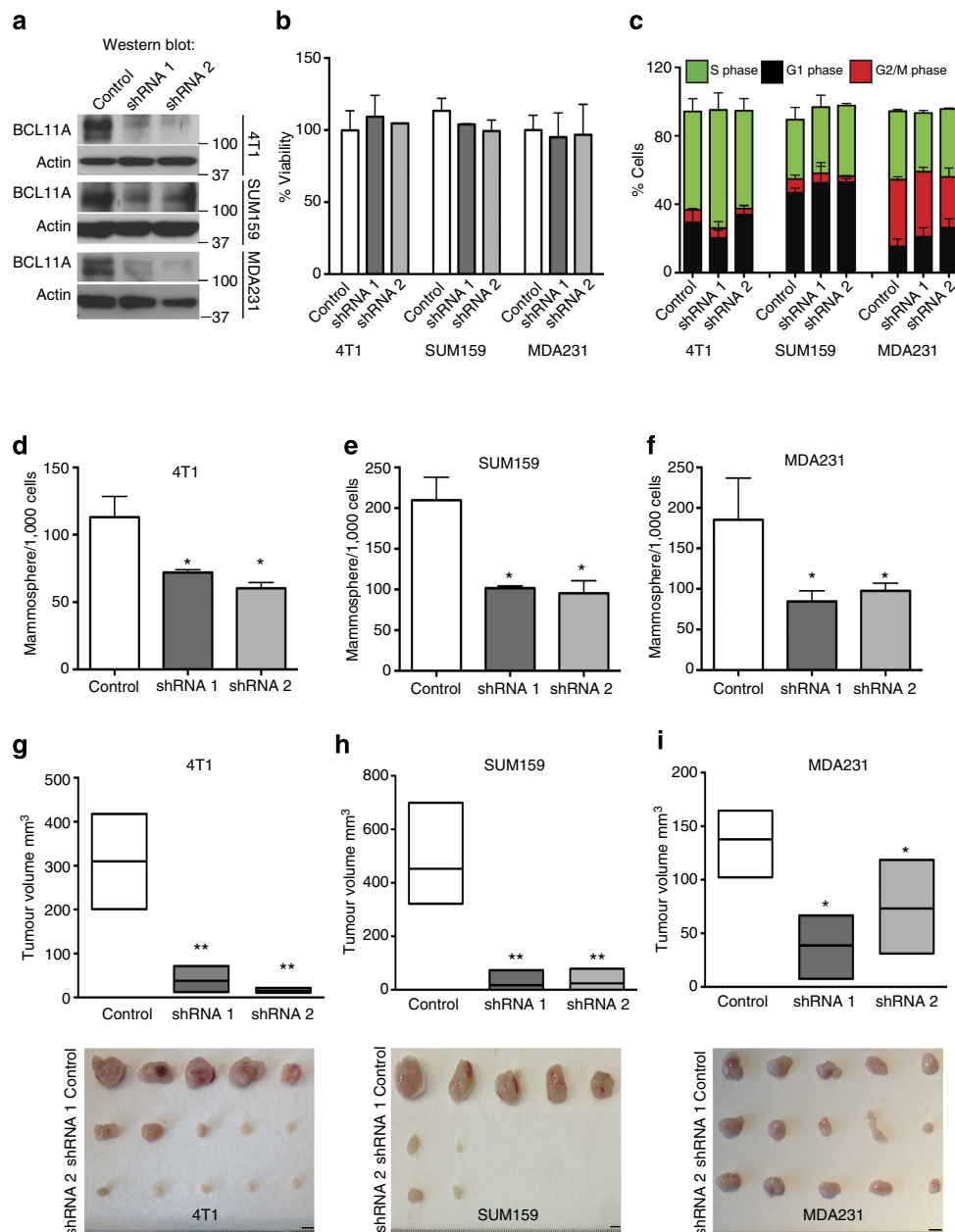
but is undetectable in any of the luminal cell lines tested (Supplementary Fig. 5a). Next, we assessed if disrupting *BCL11A* expression could affect the clonogenic and oncogenic potential of the TNBC cell lines. To inactivate *BCL11A* in these cells, we performed shRNA knockdown experiments (Supplementary Fig. 5b) in the TNBC cell lines 4T1 (mouse), MDA231, SUM159 and HMLER (human). Knockdown of *BCL11A* had no significant impact on cell viability, cell cycle kinetics or cell death (Fig. 4a–c and Supplementary Fig. 5b,d). However, *BCL11A* knockdown significantly reduced the clonogenic capacity of all four cell lines (Fig. 4d–f and Supplementary Fig. 5c). To assess tumourigenic potential, *BCL11A* knockdown cells were injected subcutaneously into NSG recipients. Robust tumours developed from the control 4T1, MDA231, SUM159 and HMLER cells within 25 days. In contrast, the *BCL11A* knockdown cells produced tumours of significantly reduced sizes (Fig. 4g–i and Supplementary Fig. 5c). Furthermore, primary and secondary limiting dilution transplantations of MDA231 control or shRNA1 cells revealed a reduction in the number of tumour-initiating cells during the secondary transplants from 1/123 to 1/667 (Supplementary Fig. 5e).

***Bcl11a* is required for the development of DMBA tumours.** To examine the role of *BCL11A* in mammary tumour development *in vivo*, we generated *Bcl11a* conditional knockout (cko) mice (referred to as *flox/flox*; Supplementary Fig. 6a), as germline deletion of *Bcl11a* causes neonatal lethality<sup>27</sup> and crossed them to the inducible *Rosa26-CreERT2*. As a tumour model, we used the potent carcinogen DMBA (7,12-dimethylbenz(a)anthracene) in combination with medroxyprogesterone acetate (MPA) to promote TNBC-like tumours in the mouse<sup>28,29</sup>. To minimize the effects of *Bcl11a* deletion on non-mammary tissues, we



**Figure 3 | High levels of *BCL11A* enhance clonogenicity of mammary epithelial cells and promote tumourigenesis.** (a) Comparison of colony numbers in matrigel from Eph4-11A cells or from the control cells. Data are presented as mean  $\pm$  s.d. ( $n=3$ ). Image on the right are depicting Eph4-control and Eph4-11A mammospheres grown in Matrigel (scale bar, 100  $\mu$ m).  $P$  value indicates student's  $t$ -test. (b) Comparison of the number of floating mammospheres formed from human HMLE-11A cells or the control cells. Data are presented as mean  $\pm$  s.d. ( $n=3$ ). Images on the right are of floating mammospheres formed by HMLE-control and HMLE-11A-expressing cell (scale bar, 200  $\mu$ m).  $P$  value indicates student's  $t$ -test. (c) Graph depicting the size difference between tumours at 6 weeks after injection of Eph4-control and Eph4-11A cells orthotopically into contralateral mammary fat pads. Data are presented as mean  $\pm$  s.d. ( $n=6$ ).  $P$  value indicates student's  $t$ -test. (d) Kaplan-Meier survival curve depicting the percentage of tumour-free mice injected with either HMLE-control or HMLE-11A cells ( $n=4$ ). (e) Unsupervised clustering of the HMLE-11A tumours in the mouse with human tumours from the METABRIC study based on the PAM50 (ref. 3) gene expression. Nearest centroid correlation score is plotted against the various subtypes for all three tumours.



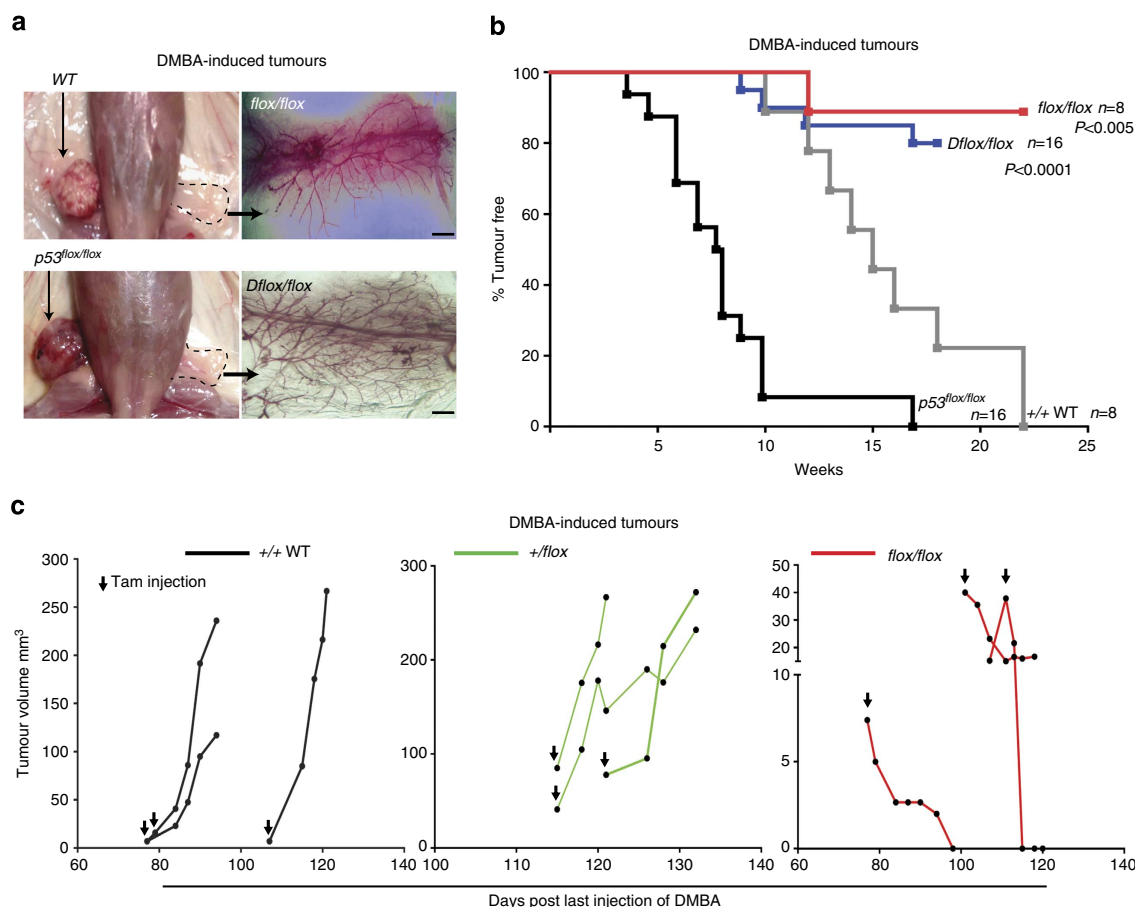


**Figure 4 | *BCL11A* knockdown in TNBC cells reduces tumour development.** (a) Western blot showing *BCL11A* knockdown efficiency in 4T1, SUM159 and MDA231 cells transfected with control (scramble), shRNA1 or shRNA2 vectors. (b) MTS cell viability assay (see Methods) shows that *BCL11A* knockdown does not affect cell viability in culture. (c) EdU cell cycle analysis showing that *Bcl11a* knockdown does not affect cell cycle kinetics in all tested cell lines. Data are presented as mean  $\pm$  s.d. (n = 3). (d-f) Comparison of colony numbers from control, shRNA1 and shRNA2 in 4T1, SUM159 and MDA231 cells. Data are presented as mean  $\pm$  s.d. (n = 3). *P* value indicates student's *t*-test. (g-i) Graph depicting the reduction in tumour size observed when shRNA1 or shRNA 2 transfected 4T1, SUM159 or MDA231 cells are injected subcutaneously into mice compared with control. Image under graph shows the actual tumours measured (scale bar, 5 mm). Unpaired *T*-test was performed on d-i and \* indicates  $P < 0.05$  and \*\* indicates  $P < 0.005$ .

transplanted mammary tissue from 8- to 12-week-old control (wild type) or *flox/flox* virgin female mice into contralateral cleared fat pads of female NSG mice followed by DMBA mutagenesis as illustrated in Supplementary Fig. 6b. By week 15, after the last dose of DMBA was administered, palpable tumours were visible in the mammary glands engrafted with the control mammary cells, but not with the *flox/flox* cells (Fig. 5a). By week 22 post DMBA treatment, all control cell engraftments (8/8) developed tumours compared with only one from *flox/flox* mammary cells (1/8) (Fig. 5b). qRT-PCR analysis of this tumour revealed expression of *Bcl11a* probably owing to incomplete

Cre-*loxP* recombination (Supplementary Fig. 6c, sample T1). Also, qRT-PCR and IHC results revealed that tumours upregulated *Bcl11a* expression in response to DMBA-induced carcinogenesis (Supplementary Fig. 6c-d). These data thus reveal a requirement for *Bcl11a* in DMBA-induced mammary tumourigenesis.

To investigate *Bcl11a* oncogenic activity in the DMBA model further, we performed the DMBA mutagenesis experiment using *Trp53flox/flox*<sup>30</sup> (*p53* single cko) or *Bcl11aflox/flox/p53flox/flox* (cko alleles for both *p53* and *Bcl11a* or *Dflox/flox*) mammary tissues. In the recipients transplanted with *Trp53flox/flox* cells,



**Figure 5 | *Bcl11a* is required in DMBA-mediated tumorigenesis.** (a) Mouse images and fat pads whole-mount fat pads of either WT, *flox/flox*, *p53<sup>flox/flox</sup>* or *Dflox/flox* mammary cells. (scale bar, 500  $\mu$ m) (b) Quantification of the tumours with DMBA-mediated tumour development from the four depicted groups of engrafted cells over a period of 26 weeks. A log-rank (Mantel-Cox) test was used to compare the two groups and calculate the *P* value. (c) Tumour size quantification of DMBA-mediated tumours in WT, +/-flox and *flox/flox* mice. Mice were checked regularly for tumour development and once detected, Cre activation was induced using three injections of tamoxifen (first day of injection is indicated by black arrow). Tumours size was then monitored for up to 20 days or until they reach a critical size.

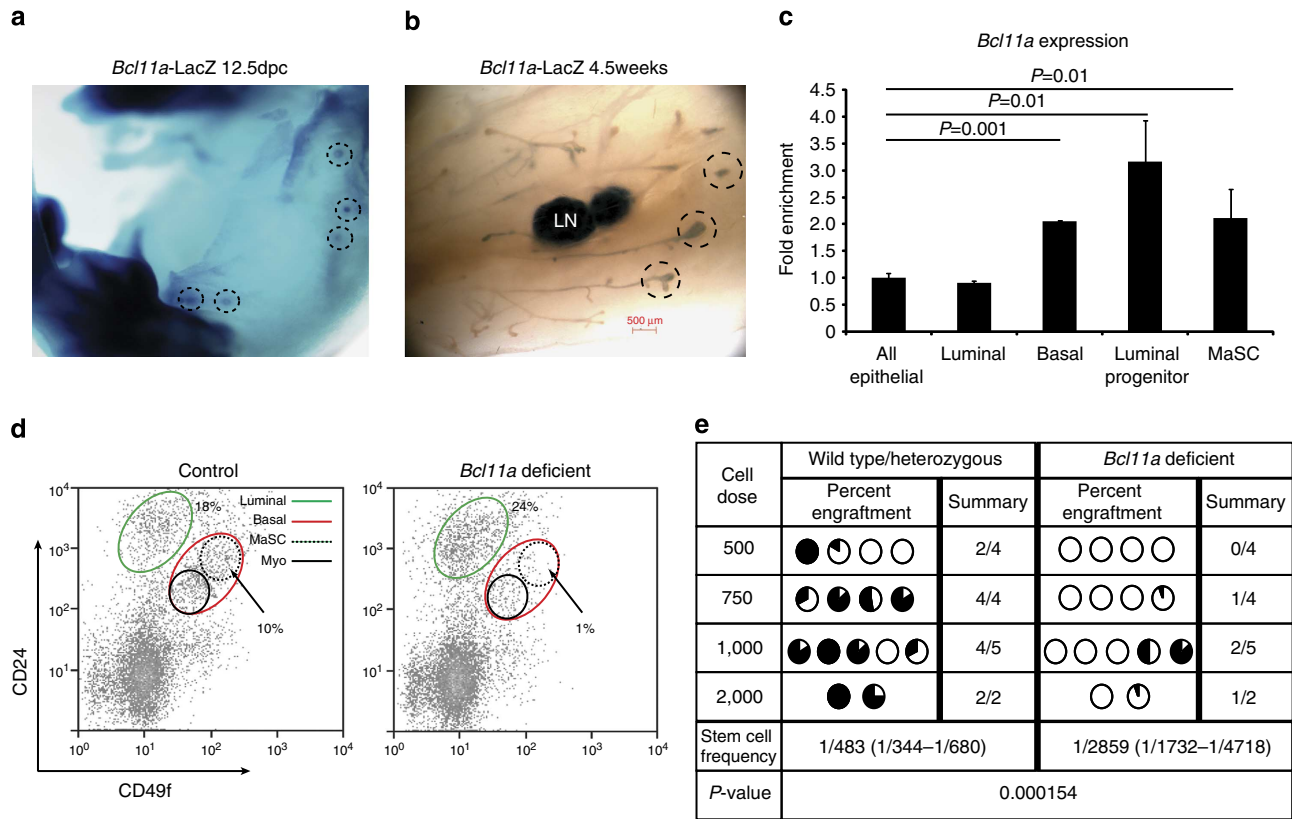
palpable tumours were detectable as early as 4 weeks after the last injection of DMBA, and most tumours were detectable by week 10 (Fig. 5b; *n* = 16). However, deletion of *Bcl11a* together with *p53* in *Dflox/flox* mice severely delayed tumour development with only 4 out of 16 mice developing tumours by week 17 (Fig. 5b). This result indicates that BCL11A is a potent oncogene and is required in concert with *p53* for tumour development.

***Bcl11a* is required for the maintenance of DMBA tumours.** Although *Bcl11a* is important for DMBA-induced mammary tumour formation, it is more clinically relevant if it has functions in mammary tumour progression and maintenance. We thus performed the DMBA mutagenesis on WT, *flox/+* and *flox/flox* mammary epithelial cells before the induction of *Bcl11a* deletion. Only when mammary tumours were detected and measured, the mice were then injected with tamoxifen to induce *Bcl11a* deletion. As shown in Fig. 5c, deletion of *Bcl11a* caused a significant reduction in tumour size as soon as 5 days post deletion. On contrary, tumours from the control heterozygous donor cells continued to grow post tamoxifen injection (Fig. 5c). The requirement of *Bcl11a* in the established mouse mammary tumours is consistent with the decreased tumorigenesis of BCL11A knockdown breast cancer cells and underscores its candidature for therapeutic development.

### ***Bcl11a* is required for mammary stem and progenitor cells.**

To understand the biological function of *Bcl11a* in healthy mammary epithelial cells, we generated a *Bcl11a-lacZ* knock-in mouse to determine the temporal and spatial expression of *Bcl11a* in the mammary gland (Supplementary Fig. 7a). X-gal staining of the reporter embryos revealed that *Bcl11a* was expressed in the mammary placodes from 12.5dpc (Fig. 6a). At puberty, *Bcl11a* was expressed in the cap cells of the terminal end buds, a region thought to harbour stem cells<sup>31</sup> (Fig. 6b). During adult mammary gland development, *Bcl11a* exhibited a dynamic expression pattern with a marked increase at early gestation and a gradual decline towards lactation and involution (Supplementary Fig. 7b). qRT-PCR analysis of RNA samples from several mammary epithelial compartments<sup>32,33</sup> detected higher levels of *Bcl11a* expression in the luminal progenitors (CD49b<sup>+</sup>/CD24<sup>hi</sup>), the basal cells (CD49F<sup>hi</sup>/CD24<sup>+</sup>) and the mammary stem cell (MaSC) (CD49F<sup>hi</sup>/CD24<sup>med</sup>)-enriched population (Fig. 6c).

We next induced *Bcl11a* deletion and analysed the mammary epithelial fluorescence-activated cell sorting profile 3 weeks post deletion. The basal mammary epithelial cells from the *flox/flox* mice appeared to be depleted, and in particular the MaSC fraction (Fig. 6d). In addition, *Bcl11a* deletion caused a significant decrease in the number of luminal colony-forming cells (CFCs) (Supplementary Fig. 7c). To functionally demonstrate loss of



**Figure 6 | Expression and critical roles of *Bcl11a* in mouse MaSCs and progenitors.** (a,b) X-gal staining of a 12.5dpc embryo and whole mount of the mammary gland from a 5-week-old *Bcl11a*-LacZ/+ female virgin mouse. The dashed circles highlight the mammary placodes and the terminal end buds. LN: lymph node. (scale bar, 500  $\mu$ m) (c) qRT-PCR for *Bcl11a* in different mammary epithelial cell compartments that were fluorescence-activated cell sorting-purified using antibodies for CD24, CD49f and CD49b. Data are presented as mean  $\pm$  s.d. ( $n = 3$ ) and  $t$ -test was performed and the  $P$  values are displayed on the plot. (d) Depletion of the MaSCs-enriched population (CD24<sup>med</sup>CD49f<sup>hi</sup>, dashed lines) in the *Bcl11a*-deficient mammary gland detected by flow cytometric analysis. (e) Limiting dilution transplant (fat pad) assay showing severely compromised engraftment of *Bcl11a*-deficient MaSCs. Stem cell frequency calculation is described in the Methods.

MaSC activities upon *Bcl11a* deletion and to determine that the defects are cell-autonomous, we transplanted control and *flox/flox* cells at limiting dilution into cleared fat pads of NSG mice (see Methods). We found approximately sixfold reduction in stem cell frequency from 1/483 to 1/2859, in the *Bcl11a*-deficient mammary gland (Fig. 6e). Reduction of MaSCs and progenitors in the *Bcl11a*-deficient mammary gland was also reflected in the altered expression of the MaSC gene expression signature<sup>34</sup> (Supplementary Table 1) (Supplementary Fig. 7e).

Discussion

We have demonstrated here that the transcription regulator BCL11A is a novel breast cancer gene. By investigating cancer genomics data from ~3,000 patients (METABRIC and TCGA), BCL11A was significantly expressed at higher levels in TNBC and particularly in BLBC/IC10 tumours both at RNA and protein levels. Experimentally, we have shown that disrupting BCL11A expression in TNBC cell lines and in the mouse significantly reduced tumour development and maintenance. At the cellular level, *Bcl11a* is expressed and required in both MaSCs and luminal progenitor cells in the mammary gland. Lineage tracing experiments in the future will determine if *Bcl11a* is expressed in the recently identified lineage-restricted luminal and basal progenitor cells<sup>35</sup> or in the bipotent MaSCs<sup>36</sup>. Importantly, given the recent implication of luminal progenitors as the ‘cell of origin’ of BLBC<sup>37,38</sup>, it will be important to ascertain if *Bcl11a*

upregulation in luminal progenitor cells is one of the earliest steps in TNBC development.

In addition, it will be important to identify how BCL11A is transcriptionally regulated and what are its downstream targets in TNBC. In erythrocytes, KLF1 has been shown to affect BCL11A expression<sup>39</sup>, while in non-small cell lung cancer MIR30A has been suggested to regulate BCL11A expression<sup>40</sup>. We found no correlation between KLF1 or MIR30A and BCL11A expression in the TCGA data set (Supplementary Fig. 8), suggesting that BCL11A regulation could be context dependent. In terms of downstream targets, in leukaemia, it has been shown that BCL11A abrogates p21 transcription possibly via direct regulation of SIRT1 (refs 41,42). Previous work from our lab also showed that in B cells, BCL11A induces MDM2 expression, which is a negative regulator of p53 (ref. 43). However, the TCGA data does not indicate a strong correlation between BCL11A and SIRT1 or MDM2 expression at least in the tumour context (Supplementary Fig. 8). Therefore, identifying the putative BCL11A regulators and its downstream targets in the breast epithelial cells should clarify its molecular and cellular roles in TNBC.

In conclusion, through cancer genomics, *in vitro* assays, experimental xenograft models and mouse genetics, we have demonstrated in this study that BCL11A is a new breast cancer gene and a critical regulator in normal mammary epithelial development. These results warrant further investigation of BCL11A as a potential candidate for TNBC-targeted therapy.



## Methods

**Mouse strains and breeding.** All experimental animal work was performed in accordance to the Animals (Scientific Procedures) Act 1986, UK and approved by the Ethics Committee at the Sanger Institute. *Bcl11a* bacterial artificial chromosomes (BACs) were identified from the 129/SvJ mouse BAC library (Sanger Institute) and used to generate the *Bcl11a-lacZ*- and *Bcl11a* cko-targeting vectors. For *Bcl11a-lacZ* reporter, targeting construct (Supplementary Fig. 7a) was generated based on the recently published strategy<sup>44</sup>. For the *Bcl11a* cko mouse, targeting construct (Supplementary Fig. 6a) was generated based on the original recombineering strategy<sup>45</sup>. Gene targeting in embryonic stem (ES) cells and chimera production were performed according to the standard procedures. The *Bcl11a* cko line was then crossed to the Rosa26-Cre-ERT2 mouse line described previously<sup>46</sup>. The *p53* cko line was described previously<sup>30</sup>. Homozygous *p53* cko mice were crossed to the *Bcl11a*/Cre-ERT2 line described above and the F1 generation was mated to generate mice doubly conditional for *Bcl11a* and *p53*. Genotyping was confirmed using the primers listed in Supplementary Table 2. Cre activation was mediated by three injections of 1 mg tamoxifen per mouse over 3 days.

**Mammary epithelial cell isolation and analysis.** Mammary epithelial cells were dissociated using a mixture of collagenase (Roche) and hyaluronidase (Sigma), and cells were stained using the following primary antibodies: biotinylated anti-CD45 (clone 30-F11; eBioscience, 1:500), anti-Ter119 (clone Ter119; eBioscience, 1:500) and anti-CD31 (clone 390; eBioscience, 1:500); anti-CD24-R-phycoerythrin (PE; clone M1/69, eBioscience, 1:500), anti-CD49f-Alexa Fluor 647 (AF647; clone GoH3, eBioscience, 1:100), anti-CD49b-Alexa Fluor 488 (AF488; clone Hma2; eBioscience, 1:500) and Scal-Alexa Fluor 647 (AF647; clone D7, eBioscience, 1:500). Secondary antibodies used: Streptavidin-PE-Texas Red (PE-TR; Molecular Probes, 1:500). Apoptotic cells were excluded by elimination of propidium iodide-positive cells. Flow cytometric analysis was done using CyAn ADP (DakoCytomation) and all sorts were performed using MoFlo (DakoCytomation) and gates were set to exclude >99.9% of cells labelled with isoform-matched control antibodies conjugated with the corresponding fluorochromes. For whole-mount analysis, abdominal glands (no. 4) were spread out using forceps on a glass slide and incubated in Carnoy's fixative overnight. The slide was washed in water and placed in carmine alum (Sigma) stain overnight. The slide was again washed with ethanol and cleared in Xylene for 1 day before documentation. For histological analysis, abdominal glands were fixed in 4% formaldehyde in PBS for 24 h at room temperature. The glands were transferred to 70% ethanol and stored at  $-20^{\circ}\text{C}$  until embedding and sectioning. All tissues were embedded in wax and sectioned at 5  $\mu\text{m}$  before being stained with haematoxylin and eosin.

**Mammary CFC assay.** For colony-forming assays, the medium used was (human) NeuroCult NS-A Proliferation Medium (StemCell) supplemented with 5% fetal bovine serum, 10 ng ml<sup>-1</sup> epidermal growth factor (Sigma), 10 ng ml<sup>-1</sup> basic fibroblast growth factor (Peprotech) and N2 Supplement (Invitrogen); the cultures were maintained at 37 °C/5% CO<sub>2</sub> for 7 days; then fixed using ice-cold acetone/methanol (1:1) and visualized using Giemsa staining (Merck). Lin<sup>-</sup> CD24<sup>hi</sup>CD49b<sup>+</sup> luminal progenitors from the *flox/flox* mammary gland were sorted and plated with irradiated feeders in colony-forming assay medium for 6 days before the number of mammary CFCs was enumerated.

**Transplantation of mammary epithelium.** Mammary epithelial cells (basal fraction) from tamoxifen-injected and non-injected *flox/flox* or *flox/+* mice were sorted based on CD24/CD49f and transplanted in limiting doses (500/750/1,000/2,000 cells) into cleared fat pads of 3-week-old NSG females. In each case, non-injected and tamoxifen-injected epithelial cells were engrafted into contralateral glands of the same recipient mice. The recipient mice were impregnated 3–6 weeks after transplant and outgrowths produced were dissected, stained with carmine and scored. Stem cell frequency was calculated using L-Calc (StemCell Technologies).

**DMBA/MPA tumorigenesis protocol.** Mammary fragments were transplanted into cleared fat pads of 3-week-old NSG mice. At the time of surgery, the MPA slow release pellet (Innovative Research of America) was also implanted subcutaneously. The mice were allowed to recover for 2 weeks and then *Bcl11a* deletion was induced using three injections of tamoxifen. One week after deletion of *Bcl11a*, 1 mg of DMBA (Sigma) was administered orally; this was followed by three further doses of 1 mg of DMBA over 3 weeks. Mice were then examined weekly for tumour incidence and killed when tumours reached the legal limit.

**Transfection and mammosphere assays.** EpH4 (gift from Professor Christine Watson) and MDA231 (ATCC) cells were cultured to confluence in 1:1 DMEM:F12 (Invitrogen) media containing 10% fetal calf serum (FCS; FetalcloneIII, Clontech). 4T1 (ATCC) cells were cultured in Roswell Park Memorial Institute (RPMI) media (Invitrogen) containing 10% FCS, and SUM159 cells (gift from Dr Charlotte Kuperwasser) were cultured in Ham's F12 (Sigma), 5% FCS, insulin (5  $\mu\text{g ml}^{-1}$ , Sigma) hydrocortisone (1  $\mu\text{g ml}^{-1}$ , Sigma) and 1  $\times$  Penicillin Streptomycin Glutamine (PSG) (Gibco). HMLE and HMLER cells (gift from Professor Robert Weinberg)

were cultured in complete HuMEC media (Invitrogen). The control or the *Bcl11a* overexpression piggyBac vectors were delivered into cells using the Amaxa Basic Nucleofactor Kit for primary mammalian epithelial cells (Lonza) according to the manufacturer's recommendations. Transfected cells were maintained at 37 °C/5% CO<sub>2</sub> for 48 h. Cells were then cultured in puromycin (1–5  $\mu\text{g ml}^{-1}$ ) for 48 h to allow for selection. To induce BCL11A expression in EpH4 and HMLE cells, doxycycline (Clontech) was used at a final concentration of 1.0  $\mu\text{g ml}^{-1}$ . Floating or Matrigel-embedded mammosphere were cultured and passaged as previously described<sup>47</sup> in ultra-low attachment plates (Corning).

**RNA knockdown.** *BCL11A* shRNA sequences were obtained from the TRC consortium<sup>48</sup> (TRCN0000033449 and TRCN0000033453) were cloned into piggyBac transposon vector (PB-H1-shRNA-GFP). 4T1, SUM159, MDA231 and HMLER cells were transfected with 4.0  $\mu\text{g}$  of piggyBac vector using Amaxa Basic Nucleofactor Kit for primary mammalian epithelial cells (Lonza) and GFP<sup>+</sup> cells were sorted/analysed 24–48 h later.

**RNA extraction and real-time PCR analysis.** RNA from sorted cells was extracted using PicoPure RNA isolation kit (Molecular Devices) according to the manufacturer's instructions. RNA from mammary tissue and cell lines was extracted using Tri-Reagent (Invitrogen) according to the manufacturer's instructions. Complementary DNA (cDNA) was synthesized from 1 to 2  $\mu\text{g}$  of total RNA using the Transcriptor Reverse Transcription cDNA Synthesis Kit (Roche). RT-PCR was performed using Hi-Fidelity Extensor mix (Thermo) using primers listed in Supplementary Table 2. Quantitative real-time PCR detection of cDNA was performed using SYBR Green Master Mix (Sigma, ABI and Invitrogen) according to supplier's recommendations. The real-time PCR reactions were run in ABI-7900HT (Applied Biosystems) in triplicate. Primers used for real-time PCR on mouse samples were designed using PrimerBank<sup>49</sup> website (<http://pga.mgh.harvard.edu/primerbank/>) and listed in Supplementary Table 2. All primers were purchased from Sigma-Aldrich. For real-time PCR on human samples Taqman gene expression probes (Life Technologies) were used.

**Cell cycle analysis.** A total of 150,000 control or *BCL11A* knockdown cells were seeded in six-well plates and allowed to recover for 48 h. Cells were then incubated with 5  $\mu\text{M}$  Edu (Invitrogen) for 1 h. Cells were fixed and assayed using the Edu flow cytometry detection kit (Invitrogen) following the manufacturer's instructions.

**Annexin v assays.** A total of 100,000 control or *BCL11A* knockdown cells (in triplicates) were seeded in six-well plates and allowed to recover for 48 h. Cell were then collected and stained using the Annexin-V-AF647 (BioLegend) following the manufacturer's instructions, and cells were then quantified using fluorescence-activated cell sorting.

**Cell viability assay.** A total of 1,000 control or *BCL11A* knockdown cells (in triplicates) were seeded in 96-well plates and allowed to recover for 48 h. Cells were then incubated with CellTiter Aqueous One Solution (Promega) for 4 h following the manufacturer's instructions. Absorbance was then measured at 490 nm using a plate reader (Bio-Rad).

**Western blotting and IHC.** Protein samples were prepared as described previously<sup>14</sup> and probed using anti-Bcl11a (Abcam, Clone 14B5, 1:1000) and Actin (Cell Signalling, 1:10000). For IHC analysis, BCL11A (Abcam (14B5, 1:50); CK14 (Abcam; 1:100) and ER $\alpha$  (SCBT; 1:50) were used. Staining was detected using AF488- or Cy3-conjugated secondary (Sigma) and bisbenzimidazole-Hoechst 33342 (Sigma). Fluorescence microscopy was carried out using a Zeiss Axiophot microscope equipped with a Hamamatsu Orca 285 camera, with images visualized, captured and manipulated using Simple PCI 6 (C Imaging Systems). The haematoxylin- and eosin-stained samples were visualized on a LEICA light microscope, while the mouse mammary gland whole mounts were visualized using the LEICA MZ75 light microscope.

**Microarray analysis.** The intensity value for each probe set was calculated and the average of each gene was computed before the data analysis. For the quality control (QC) step, a set of intensity value of control genes were examined. All data were normalized and scaled by Partek Genomic Suite 6.4. Principal components analysis was performed to show the distribution of samples, eliminating outliers. Differentially expressed genes were selected by one-way analysis of variance by the factor of KO versus wild type,  $P$  value < 0.08. Hierarchical clustering of selected genes was performed to show the expression pattern. The resulting genes then underwent a pathway analysis (GeneGO: <http://www.genego.com>) to determine the biological significance of the data.

**Xenograft tumorigenesis assays.** One hundred thousand EpH4, HMLE, 4T1, MDA231, SUM159 or HMLER cells were suspended in 25% Matrigel (BD

Biosciences) and HBSS, and injected into either cleared contralateral number 4 mammary fat pads of 3-week-old female mice or subcutaneously in 6–12-week-old female NSG mice. For secondary transplants, tumours were dissociated using collagenase/hyaluronase mix (Roche) for 16 h and viable cells were counted and injected into NSG recipient mice at the indicated doses.

**METABRIC analysis.** Matched DNA and RNA were extracted for tumours. CN analysis was performed using the Affymetrix SNP 6.0 platform. The arrays were pre-processed and normalized using CRMAv2 (ref. 50) method from aroma.affymetrix. Allelic-crosstalk calibration, probe sequence effects normalization, probe-level summarization and PCR fragment length normalization were performed for each array. The intensities obtained were normalized against a pool of 473 normals for the samples with no matched pair or against their matched normal when available (258 samples). The log-ratios were then segmented using the circular binary segmentation algorithm<sup>51</sup> in the DNACopy Bioconductor package<sup>52</sup>. Then, callings into five groups (homozygous deletion, heterozygous deletion, neutral CN, gain ( $>2$ ) and amplification ( $>3$ )) were made using thresholds based on the variability of each sample and their proportion of tumoural cells. RNA analysis was performed using Illumina HT-12 v3 platform and analysed using beadarray package<sup>53</sup>. BASH<sup>54</sup> was used to correct for spatial artifacts. The bead-level data were summarized and a selection of suitable probes based on their quality was done using the re-annotation of the Illumina HT-12v3 platform<sup>55</sup>. The samples were classified into the five breast cancer subtypes using PAM50 (ref. 56), but only those genes with a probe with perfect annotation on the chip were considered. A mixture model was used to classify BCL11A expression into low and high values<sup>57,58</sup>.

**TCGA data analysis.** All TCGA data and figures were accessed, analysed and generated using the cBio Cancer Genomics Portal<sup>59</sup>. All data included in this manuscript is in agreement with the TCGA publication guidelines.

**METABRIC IHC analysis.** A subset of patients enrolled in the METABRIC study with tumour samples represented in tissue microarrays (TMAs) were included for the detection of BCL11A protein expression by IHC. TMAs were constructed from formalin-fixed paraffin-embedded tumour blocks as previously described<sup>60</sup>. Each tumour was represented by a single 0.6-mm tissue core. A total of 439 tumours were included arising from 436 patients (three were synchronous tumours arising in the contralateral breast). CN and gene expression data was available for 368 of these tumours for correlative analyses. Three micrometre TMA sections were dewaxed in xylene and rehydrated through graded alcohols. IHC was conducted using a BondMax Autoimmunostainer (Leica, Bucks, UK). Antigen retrieval was achieved by heating TMA sections in pH 6 citrate buffer for 20 min. Primary mouse monoclonal (clone 14B5) antibody bound to BCL11A (ab19487, AbCam) was diluted to 1:200 and detected using a BOND Polymer detection kit (Leica) and signal developed with 3,3'-diaminobenzidine (DAB). Stained TMA sections were digitized using the Ariol (Genetix Ltd, Hampshire, UK) platform for scoring by a pathologist (H.R.A.). The ordinal Allred scoring system was used for assessing the amount of staining present in tumour cells accounting for the intensity (0 = no staining, 1 = weak, 2 = moderate and 3 = strong) and proportion (0 = 0%, 1 =  $<1\%$ , 2 = 1–10%, 3 = 11–33%, 4 = 34–66% and 5 =  $>66\%$ ) of stained cells, finally producing a summed score (intensity + proportion = Allred score) between zero and eight. Analogous to clinical practice for estrogen receptor (ER), tumours with an Allred score of  $>2$  were deemed positive for BCL11A expression and comparison with molecular subtypes was made using Pearson's  $\chi^2$  test.

**Statistical significance.** All *P* values were calculated using Student's *t*-test unless otherwise indicated in the figure legends.

## References

- Blows, F. M. *et al.* Subtyping of breast cancer by immunohistochemistry to investigate a relationship between subtype and short and long term survival: a collaborative analysis of data for 10,159 cases from 12 studies. *PLoS Med.* **7**, e1000279 (2010).
- Foulkes, W. D., Smith, I. E. & Reis-Filho, J. S. Triple-negative breast cancer. *N. Engl. J. Med.* **363**, 1938–1948 (2011).
- Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
- Carey, L., Winer, E., Viale, G., Cameron, D. & Gianni, L. Triple-negative breast cancer: disease entity or title of convenience? *Nat. Rev. Clin. Oncol.* **7**, 683–692 (2010).
- Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
- Prat, A. *et al.* Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res.* **12**, R68 (2010).
- Weigelt, B., Baehner, F. L. & Reis-Filho, J. S. The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade. *J. Pathol.* **220**, 263–280 (2010).
- TCGA. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
- Shah, S. P. *et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486**, 395–399 (2012).
- Stingl, J. & Caldas, C. Molecular heterogeneity of breast carcinomas and the cancer stem cell hypothesis. *Nat. Rev. Cancer* **7**, 791–799 (2007).
- Bouras, T. *et al.* Notch signaling regulates mammary stem cell function and luminal cell-fate commitment. *Cell Stem Cell* **3**, 429–441 (2008).
- Kouros-Mehr, H., Slorach, E. M., Sternlicht, M. D. & Werb, Z. GATA-3 maintains the differentiation of the luminal cell fate in the mammary gland. *Cell* **127**, 1041–1055 (2006).
- Asselin-Labat, M. L. *et al.* Gata-3 is an essential regulator of mammary-gland morphogenesis and luminal-cell differentiation. *Nat. Cell Biol.* **9**, 201–209 (2007).
- Khaled, W. T. *et al.* The IL-4/IL-13/Stat6 signalling pathway promotes luminal mammary epithelial cell development. *Development* **134**, 2739–2750 (2007).
- Zuo, T. *et al.* FOXP3 is an X-linked breast cancer suppressor gene and an important repressor of the HER-2/ErbB2 oncogene. *Cell* **129**, 1275–1286 (2007).
- Mehra, R. *et al.* Identification of GATA3 as a breast cancer prognostic marker by global gene expression meta-analysis. *Cancer Res.* **65**, 11259–11264 (2005).
- Desmedt, C. *et al.* Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin. Cancer Res.* **13**, 3207–3214 (2007).
- Hatzis, C. *et al.* A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. *JAMA* **305**, 1873–1881 (2011).
- Ma, X. J. *et al.* A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer Cell* **5**, 607–616 (2004).
- Schmidt, M. *et al.* The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Res.* **68**, 5405–5413 (2008).
- Weniger, M. A. *et al.* Gains of the proto-oncogene BCL11A and nuclear accumulation of BCL11A(XL) protein are frequent in primary mediastinal B-cell lymphoma. *Leukemia* **20**, 1880–1882 (2006).
- Nakamura, T. *et al.* Evi9 encodes a novel zinc finger protein that physically interacts with BCL6, a known human B-cell proto-oncogene product. *Mol. Cell Biol.* **20**, 3178–3186 (2000).
- Reichmann, E., Ball, R., Groner, B. & Friis, R. R. New mammary epithelial and fibroblastic cell clones in coculture form structures competent to differentiate functionally. *J. Cell Biol.* **108**, 1127–1138 (1989).
- Ince, T. A. *et al.* Transformation of different human breast epithelial cell types leads to distinct tumor phenotypes. *Cancer Cell* **12**, 160–170 (2007).
- Elenbaas, B. *et al.* Human breast cancer cells generated by oncogenic transformation of primary mammary epithelial cells. *Genes Dev.* **15**, 50–65 (2001).
- Ishikawa, F. *et al.* Development of functional human blood and immune systems in NOD/SCID/IL2 receptor  $\gamma$  chain(null) mice. *Blood* **106**, 1565–1573 (2005).
- Liu, P. *et al.* Bcl11a is essential for normal lymphoid development. *Nat. Immunol.* **4**, 525–532 (2003).
- Herschkowitz, J. I. *et al.* Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biol.* **8**, R76 (2007).
- Herschkowitz, J. I. *et al.* Comparative oncogenomics identifies breast tumors enriched in functional tumor-initiating cells. *Proc. Natl Acad. Sci. USA* **109**, 2778–2783 (2012).
- Jonkers, J. *et al.* Synergistic tumor suppressor activity of BRCA2 and p53 in a conditional mouse model for breast cancer. *Nat. Genet.* **29**, 418–425 (2001).
- Hennighausen, L. & Robinson, G. W. Information networks in the mammary gland. *Nat. Rev. Mol. Cell Biol.* **6**, 715–725 (2005).
- Stingl, J. *et al.* Purification and unique properties of mammary epithelial stem cells. *Nature* **439**, 993–997 (2006).
- Shackleton, M. *et al.* Generation of a functional mammary gland from a single stem cell. *Nature* **439**, 84–88 (2006).
- Lim, E. *et al.* Transcriptome analyses of mouse and human mammary cell subpopulations reveal multiple conserved genes and pathways. *Breast Cancer Res.* **12**, R21 (2010).
- Van Keymeulen, A. *et al.* Distinct stem cells contribute to mammary gland development and maintenance. *Nature* **479**, 189–193 (2011).
- Rios, A. C., Fu, N. Y., Lindeman, G. J. & Visvader, J. E. *In situ* identification of bipotent stem cells in the mammary gland. *Nature* **506**, 322–327 (2014).
- Lim, E. *et al.* Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nat. Med.* **15**, 907–913 (2009).
- Molyneux, G. *et al.* BRCA1 basal-like breast cancers originate from luminal epithelial progenitors and not from basal stem cells. *Cell Stem Cell* **7**, 403–417 (2010).

39. Zhou, D., Liu, K., Sun, C. W., Pawlik, K. M. & Townes, T. M. KLF1 regulates BCL11A expression and gamma- to beta-globin gene switching. *Nat. Genet.* **42**, 742–744 (2010).
40. Jiang, B. Y. *et al.* BCL11A overexpression predicts survival and relapse in non-small cell lung cancer and is modulated by microRNA-30a and gene amplification. *Mol. Cancer* **12**, 61 (2013).
41. Senawong, T., Peterson, V. J. & Leid, M. BCL11A-dependent recruitment of SIRT1 to a promoter template in mammalian cells results in histone deacetylation and transcriptional repression. *Arch. Biochem. Biophys.* **434**, 316–325 (2005).
42. Yin, B. *et al.* A retroviral mutagenesis screen reveals strong cooperation between Bcl11a overexpression and loss of the Nf1 tumor suppressor gene. *Blood* **113**, 1075–1085 (2008).
43. Yu, Y. *et al.* Bcl11a is essential for lymphoid development and negatively regulates p53. *J. Exp. Med.* **209**, 2467–2483 (2012).
44. Chan, W. *et al.* A recombineering based approach for high-throughput conditional knockout targeting vector construction. *Nucleic Acids Res.* **35**, e64 (2007).
45. Liu, P., Jenkins, N. A. & Copeland, N. G. A highly efficient recombineering-based method for generating conditional knockout mutations. *Genome Res.* **13**, 476–484 (2003).
46. Hameyer, D. *et al.* Toxicity of ligand-dependent Cre recombinases and generation of a conditional Cre deleter mouse allowing mosaic recombination in peripheral tissues. *Physiol. Genomics* **31**, 32–41 (2007).
47. Dontu, G. *et al.* In vitro propagation and transcriptional profiling of human mammary stem/progenitor cells. *Genes Dev.* **17**, 1253–1270 (2003).
48. Moffat, J. *et al.* A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. *Cell* **124**, 1283–1298 (2006).
49. Wang, X. & Seed, B. A PCR primer bank for quantitative gene expression analysis. *Nucleic Acids Res.* **31**, e154 (2003).
50. Bengtsson, H., Wirapati, P. & Speed, T. P. A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6. *Bioinformatics* **25**, 2149–2156 (2009).
51. Venkatraman, E. S. & Olshen, A. B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**, 657–663 (2007).
52. Seshan VE, O. A. DNACopy: DNA copy number data analysis R package <http://www.bioconductor.org/packages/2.3/bioc/html/DNACopy.html> (2010).
53. Dunning, M. J., Smith, M. L., Ritchie, M. E. & Tavaré, S. beadarray: R classes and methods for Illumina bead-based data. *Bioinformatics* **23**, 2183–2184 (2007).
54. Cairns, J. M., Dunning, M. J., Ritchie, M. E., Russell, R. & Lynch, A. G. BASH: a tool for managing BeadArray spatial artefacts. *Bioinformatics* **24**, 2921–2922 (2008).
55. Barbosa-Morais, N. L. *et al.* A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data. *Nucleic Acids Res.* **38**, e17.
56. Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27**, 1160–1167 (2009).
57. Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E. & Ruzzo, W. L. Model-based clustering and data transformations for gene expression data. *Bioinformatics* **17**, 977–987 (2001).
58. Fraley, C. & Raftery, A. E. MCLUST Version 3 for R: Normal Mixture Modeling and Model-based Clustering. Technical Report No. 504, Department of Statistics, University of Washington (2006).
59. Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
60. Kononen, J. *et al.* Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat. Med.* **4**, 844–847 (1998).

## Acknowledgements

W.T.K., S.C.L. and J.W. were funded by the Sanger Institute. W.T.K. was also funded by a BBSRC project grant and a Junior Research Fellowship, King's College, Cambridge and a Cancer Research UK Career establishment award. J.S. is funded by Cancer Research UK, The University of Cambridge and Hutchison Whampoa Limited. C.J.W. is funded by BBSRC, MRC and BCC. We would like to thank Dr Floris Fojier and Professor Allan Bradley for providing the p53 kco mice. We thank Dr Shannon Burke, Dr Sara Pensa and Lauma Skruzmane for comments on the manuscript. We thank Miss Malgorzata Gawedzka from the W.T.K. lab for the technical assistance. We thank the Sanger Institute RSF, the flow cytometry core and microarray facilities for the technical assistance. This work is supported by Wellcome Trust (Grant number—098051) (P.L.).

## Author contributions

W.T.K. performed the mouse tumour studies, the cell culture experiments and analysed the *Bcl11a* kco line. S.C.L. generated and analysed the *Bcl11a-LacZ* mouse line and the *Bcl11a* kco line. J.S. developed the flow-sorting strategies, performed some of the colony assays, designed and performed some of the fat pad transplantation experiments. X.C. performed the microarray analysis. O.M.R., H.R.A., S.-F.C., S.A. and C.C. analysed and provided all the data (CN, expression, qRT-PCR and IHC) from the METABRIC cohort. All enquiries relating to the METABRIC data set should be made directly to carlos.caldas@cancer.org.uk. J.W. assisted with mouse experiments. Y.Y. assisted with qRT-PCR and cloning of shRNA constructs. A.F. and M.S. provided the reagents. The *Bcl11a* kco allele was made by P.L. in laboratory of N.G.C. and N.A.J. C.J.W. provided the reagents and contributed to the design of the study. W.T.K., S.C.L. and P.L. designed the studies and wrote the manuscript. The overall research project and manuscript writing are supervised by P.L.

## Additional information

**Accession codes:** Microarray data have been deposited in the Gene Expression Omnibus database with the accession number GSE63386.

**Supplementary Information** accompanies this paper at <http://www.nature.com/naturecommunications>

**Competing financial interests:** The authors declare no competing financial interests.

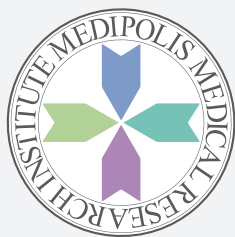
**Reprints and permission** information is available online at <http://npg.nature.com/reprintsandpermissions/>

**How to cite this article:** Khaled, W.T. *et al.* BCL11A is a triple-negative breast cancer gene with critical functions in stem and progenitor cells. *Nat. Commun.* 6:5987 doi: 10.1038/ncomms6987 (2015).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>





## Innovative proton beam therapy for breast cancer and pancreatic cancer: the Medipolis experience

### AUTHORS

Takeshi Arimura, MD<sup>1</sup>, Takashi Ogino, MD, PhD<sup>1</sup>, Ryoichi Nagata, MD, PhD<sup>2</sup>, Etsuyo Ogo, MD, PhD<sup>1,3</sup>, Yoshio Hishikawa, MD, PhD<sup>1,2</sup>

<sup>1</sup> Medipolis Proton Therapy and Research Center, Ibusuki, Japan

<sup>2</sup> Medipolis Medical Research Institute, Ibusuki, Japan

<sup>3</sup> Department of Radiation Oncology, Kurume University School of Medicine, Fukuoka, Japan

The Medipolis Proton Therapy and Research Center (MPTRC, [www.medipolis-ptrc.org/english/about.html](http://www.medipolis-ptrc.org/english/about.html)) (Figure 1) is part of Medipolis Ibusuki, a project developed jointly by the government, industry and academia in Ibusuki, a city located in Kagoshima Prefecture, Japan. The aim of Medipolis Ibusuki is to share cutting-edge medical knowledge with the rest of the world, while offering holistic treatment that aims to support every aspect of our patients' recovery in a setting that is among the best of Japan's hot spring resorts with a hotel and sports facilities.

MPTRC, the core medical facility of Medipolis Ibusuki, first offered proton beam therapy for people with cancer in 2011. It is now developing new proton beam therapies for cancers that are difficult to treat by X-ray, including pancreatic cancer<sup>1</sup> and breast cancer<sup>2</sup>. By offering this service in a resort-type setting, it aims to offer compassionate and uplifting medical care while continuously improving quality and safety.

Radiation therapy with X-rays is a valuable treatment option for many cancers but its role is limited by the risk of damage to organs adjacent to the tumour site that can sometimes be life threatening. This is because X-rays are highly penetrating, imparting ionizing energy to cells as they pass through the skin and tissues. This is not always a limiting factor: in the case of prostate cancer, for example, treatment options include radical prostatectomy, standard radiation therapy and proton therapy.

By contrast, X-ray therapy is inappropriate for early-stage breast cancer because even targeted treatment will affect important organs adjacent to the tumour but outside the targeted area, such as the heart and lungs. Early breast cancer has therefore conventionally been treated locally by surgery.

Proton beam therapy provides a means of

delivering radiation to minimize its impact outside the bounds of the tumour. Unlike X-rays, protons release most of their energy only when they stop — a phenomenon known as the Bragg peak effect. The behaviour and direction of a proton beam can be precisely controlled so that protons stop at the tumour site, minimizing the exposure of adjacent tissues to ionizing energy. At MPTRC, the treatment rooms have rotating gantries that allow the proton beam to be applied from any direction (Figure 2). It is therefore possible to treat the tumour site without affecting adjacent critical areas like the lungs and heart.

Research on proton beam therapy to treat small malignant tumours of the breast began at MPTRC in 2011. The biggest challenge was to immobilize the breast tissue so that the tumour could be targeted accurately. To date, most treatment facilities have used a specialized harness like a brassiere, but researchers at MPTRC developed a unique stabilization system called the Medipolis Technique (MPT) which maintains the breast in a fixed position. MPT was first used in the treatment of breast cancer in June 2015.

### Proton beam therapy for early-stage breast cancer

Mastectomy and axillary dissection were once the primary procedures to treat early breast cancer. Breast-conserving therapy (BCT) is now widely performed and in 2010 approximately 60% of cases of breast cancer were treated with BCT in Japan<sup>2</sup>. The rates of disease-free survival and overall survival for BCT with partial mastectomy are similar to those achieved with whole-breast irradiation or mastectomy<sup>3</sup>. Limited surgery, avoiding axillary lymph node dissection, is now more frequently offered to patients with confirmed negative sentinel lymph node biopsy (in whom tumour cells are not detected in the lymph nodes closest to the breast).



**Figure 1 | The Medipolis Proton Therapy and Research Center (MPTRC).** The center is located in the resort area of Kagoshima Prefecture. The white building (yellow circle) is MPTRC, and the brown building is the resort hotel.

After partial mastectomy, 67–86% of breast cancer recurrence occurs in the peritumoural bed region. Furthermore, ten-year survival and distant metastasis rates are higher when recurrence is located within 3 cm of the tumour bed and the secondary tumour is histopathologically identical to the primary tumour<sup>4</sup>. This suggests that partial breast irradiation (PBI) delivered to the peritumoural bed region may be sufficient to reduce post-partial mastectomy recurrence. PBI is delivered by brachytherapy, external beam irradiation and intraoperative radiotherapy.

PBI was evaluated in the RTOG 0319 trial, which reported a five-year local recurrence rate of 2.7%<sup>5</sup>. This compares favourably with outcomes after whole breast irradiation, for which the five-year local recurrence is approximately 3–5%. RTOG 0319 also showed that PBI was well tolerated, with most adverse effects being mild to moderate (grade 1, 42%; grade 2, 21%), and a low incidence of severe adverse effects (grade 3, 2%; grade 4, 0%)<sup>5</sup>.

The MPTRC is now conducting a Phase I/II clinical trial of the treatment of early breast cancer with PBI using proton beam. This study is supported by a research grant from the Japanese government's Ministry of Education, Culture, Sports, Science and Technology. Phase I will determine the optimum dose of radiation; Phase II will evaluate safety and

efficacy. The primary endpoint is safety (the frequency and severity of adverse events); based on the experience of RTOG 0319<sup>5</sup>, safety will be considered acceptable if the incidence of grade 3 or higher adverse events does not exceed 10%. Efficacy will be evaluated by secondary endpoints including time to recurrence, localization of recurrence, overall survival and cosmetic outcomes.

### Medipolis Technique: immobilization for breast cancer radiotherapy

It is not always necessary to aim for great precision when the whole breast is irradiated after breast-conserving surgery but targeted proton beam therapy requires immobilization of the breast and this presents practical difficulties for clinicians. Further, there is no consensus on the best position of the patient at fixation because the shape of the breasts may change due to the effect of gravity. Irradiation of the breast with the patient in the supine (that is, face-up) position would lead to a greater damage to the lungs and heart compared with the prone (that is, face-down) position. However, it is difficult to obtain high geometric accuracy during irradiation in the prone position because the breast is not secured and it shifts with the movement of the thorax due to respiration. The Medipolis Technique utilizes

a hybrid breast-immobilization system (HyBIS) that exploits the best aspects of the supine and prone positions.

### Using HyBIS: overview

There are three phases to HyBIS: preparation, setup and treatment. In the preparation phase, the patient is secured (except for the affected breast) on a whole body immobilization system that allows for adjustments in position. The device incorporates a photo scanning system that generates a digital image of the breast from which a unique breast cup is created. The breast cup is fitted to the breast and held in place with a specially designed retention apparatus.

With the patient and breast immobilized, the setup phase of simulation and planning treatment based on CT images can begin. These data provide the basis on which the treatment phase is implemented. Treatment ends with the removal of the breast cup and relaxation in a bathtub.

### Novel technologies

MPTRC has incorporated novel technologies in the development of HyBIS. The whole body immobilization system was designed to hold the entire body of a patient in a polycarbonate case on a purpose-built aluminum stretcher, with vacuum cushions and belts (Figure 3A).



The case can float and move vertically and horizontally by using compressed carbon dioxide gas (Figure 3B). The reason why we choose carbon dioxide gas is that a compressed gas cylinder of carbon dioxide gas can safely be brought into the magnetic resonance imaging room. It also includes a breath-synchronizing system, a localizer and a converting device to adjust the patient's position.

A converting device was developed to change or maintain the position of the patient while held in the whole body immobilization system (Figure 3C). The device, which can accommodate patients up to 1.75 m in height and up to 80 kg in weight, incorporates pressure sensors to prevent falls and a two-speed motor. The rotating angle can be adjusted within  $<0.1$  degree.

The scanning system comprises a movable array of ten cameras controlled by a touch screen terminal and powered by a large-capacity storage battery. By using battery, it is easy to handle the scanning system. This device takes a series of 10 images of the breast while the patient holds her breath. This procedure is repeated twice at slightly different locations to obtain images from 30 different viewpoints.

These 30 images generate a three-dimensional (3D) formula of the breast, from which a 3D printer creates a breast cup unique to each patient. The cup is made of resin lined with an adhesive, with a hollow center for the nipple (Figure 3D and 3E). It is connected to a fixation device with a rod (Figure 3F). The lining is highly adhesive and cannot be removed until the patient soaks herself in a bathtub. As long as the cup is dismantled and washed thoroughly, it can be used repeatedly without losing its adhesive properties.

The breast cup is fitted using a specially-designed apparatus incorporating a level gauge, a pressure sensor, three small video cameras and a localizer (Figure 3D). The apparatus can be moved and easily locked in place on the breast. Positional accuracy is ensured by a laser beam focused on a specific point of the breast and adjusted during breath-holding using the cameras (Figure 3E). The breast cup adheres to the skin with a pushing pressure and is then uncoupled from the apparatus.

The breast cup is firmly held in a device made of hard resin (Figure 3F) that enables adjustment in three axes (horizontal, longitudinal, and vertical). It is also directly linked to the whole body immobilization system to minimize the influence of the thoracic movement while breathing.



**Figure 2 | Treatment room (A) and rotating gantry (B).** The rotating gantry is a device for focusing the proton beam. The 175-ton device rotates 360 degrees around the patient, and can deliver the proton beam from any direction.

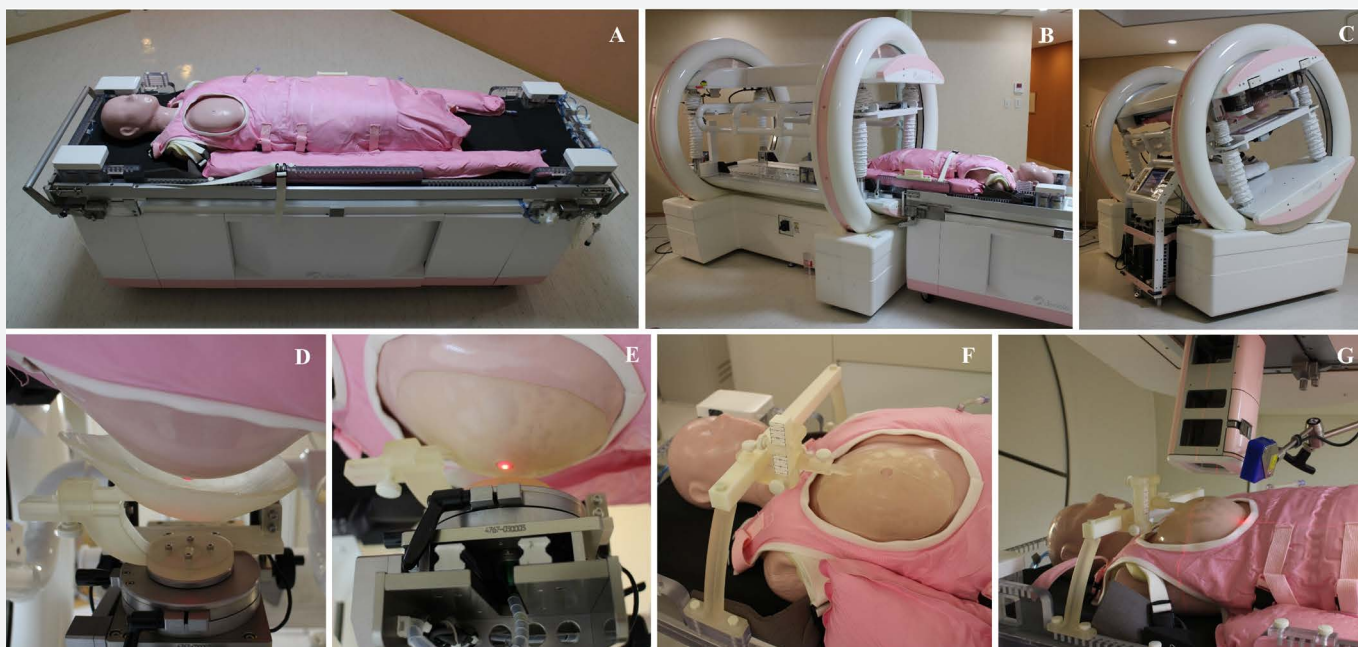
The devices used in HyBIS were tested on a specially-developed deformable dummy. The dummy was designed to simulate a 60-year-old woman and it accurately modelled the shape and quality of the breast. The dummy body was complete, from the head to the femur and was covered with thin elastic film. In addition, the dummy thorax expands and contracts under remote control to simulate respiration.

### Treatment of locally advanced pancreatic cancer

Proton beam therapy can be applied to other types of cancer where radiotherapy must

be highly targeted and its application in the treatment of pancreatic cancer, combined chemotherapy with gemcitabine, was developed at the Hyogo Ion Beam Medical Center in 2008<sup>1</sup>. Refinements such as rotating gantries and the technique of synchronizing irradiation with respiration (the respiratory gating technique), developed in the treatment of breast cancer, have led to the emergence of proton beam therapy as an important option for treating locally advanced pancreatic cancer. However, this involves a complex treatment plan and requires an experienced team.





**Figure 3 | Devices to immobilize a breast.** (A) The dummy is immobilized in the whole body immobilization system except for the right breast and the head. (B) The immobilized dummy is moved into the position converting device. (C) The position converting device is inclined at 155 degrees and the photo scanning system is set up under the gantry. (D) A breast cup is set on the fitting apparatus and a laser fixes position on the nipple. (E) The breast cup is fitted and fixed to the breast with a holding device. (F, G) An image of irradiating proton beams for breast cancer.

Effective radiotherapy of locally advanced pancreatic cancer can improve the patient's prognosis and reduce pain. This procedure, however, is challenging because the pancreas is surrounded by the stomach, duodenum and jejunum and, posteriorly, the spinal cord and aorta. The key to success is the ability to deliver a radical dose of radiation to the pancreatic tumour while limiting the exposure of the surrounding digestive tract.

X-ray radiotherapy for pancreatic cancer is associated with a high risk of adverse effects, even with mitigating strategies such as stereotactic body radiation therapy or intensity-modulated radiation therapy. This leads to an interruption of therapy, or a dose reduction or even discontinuation of treatment. Conversely, the Bragg peak effect of a proton beam means that, when therapy is delivered from the patient's back, the digestive tract is spared.

Hyogo Ion Beam Medical Center uses the field-within-a-field technique in the treatment of pancreatic cancer. This entails establishing two target areas for irradiation: a large area involving the adjacent digestive tract and a small area solely of the tumour. In this way, proton beam therapy can be targeted so that the tumour receives a higher dose of radiation than the surrounding normal tissues, increasing the effectiveness and reducing

the toxicity of treatment. After completion of proton beam therapy, patients then receive adjuvant chemotherapy.

Dr Yoshio Hishikawa began treating locally advanced pancreatic cancer at the Hyogo Ion Beam Medical Center<sup>1</sup> and he has further developed his work at MPTRC to optimize the dose of radiation to individual patients. High-level team performance by physicians specializing in radiation therapy, medical physicists, and radiation technologists, supported by nurses, is required to deliver this service effectively. As of August 2015, more than 140 patients with pancreatic cancer have been successfully treated at MPTRC.

### Reaching out

MPTRC has now treated more than 1,500 cancer patients, most of whom are Japanese. Recently, however, the Center has begun to admit more people from overseas, including China and Russia. In September 2013, MPTRC was the first proton therapy facility in Japan to be accredited by Joint Commission International ([www.jointcommissioninternational.org](http://www.jointcommissioninternational.org)). This recognition of the quality and safety of the services at MPTRC supports the center's aim of offering patient-friendly proton beam therapy and traditional Japanese hospitality, or *Omotenashi*, to people with cancer worldwide.

### REFERENCES

1. Terashima, K., Demizu, Y., Hashimoto, N. *et al.* A phase I/II study of gemcitabine-concurrent proton radiotherapy for locally advanced pancreatic cancer without distant metastasis. *Radiother. Oncol.* **103**, 25–31 (2012).
2. Japan Breast Cancer Society. Investigative report on registration of breast cancer patients in Japan. ([http://www.jbcs.gr.jp/english\\_new/REPORT/REPORT.pdf](http://www.jbcs.gr.jp/english_new/REPORT/REPORT.pdf)).
3. Early Breast Cancer Trialists' Collaborative Group. Effects of radiotherapy and of differences in the extent of surgery for early breast cancer on local recurrence and 15-year survival: an overview of the randomised trials. *Lancet* **366**, 2087–2106 (2005).
4. Huang, E., Buchholz, T.A., Meric, F. *et al.* Classifying local disease recurrences after breast conservation therapy based on location and histology. *Cancer* **95**, 2059–2067 (2002).
5. Vicini, F., Winter, K., Straube, W. *et al.* A phase I/II trial to evaluate three-dimensional conformal radiation therapy confined to the region of the lumpectomy cavity for stage I/II breast carcinoma: Initial report of feasibility and reproducibility of Radiation Therapy Oncology Group (RTOG) Study 0319. *Int. J. Radiat. Oncol. Biol. Phys.* **63**, 1531–1537 (2005).

# Brain disorders across the lifespan

Research to achieve nervous system health worldwide





FOREWORD **OPEN**

## Brain disorders across the lifespan

*Nature* 527, S150 (19 November 2015), DOI: 10.1038/nature16027

This article has not been written or reviewed by *Nature* editors. *Nature* accepts no responsibility for the accuracy of the information provided.

**T**he breadth and complexity of brain and other nervous-system disorders make them some of the most difficult conditions to diagnose and treat, especially in the developing world where there may only be one psychiatrist or neurologist in an entire country. These disorders occur throughout the lifespan — from infants starved of oxygen during difficult births, to children whose cognitive ability is stunted owing to malnutrition or exposure to infections or toxins, and to adults who develop depression, movement disorders or dementia. Mental health and behavioural issues are the cause of the largest burden of disability, and account for a staggering 184 million disability-adjusted life years (DALYs)<sup>1</sup>, according to the US Institute of Health Metrics and Evaluation.

That is why the Fogarty International Center (Fogarty) has been working with its National Institutes of Health (NIH) partners for more than 10 years through the *Brain Disorders in the Developing World: Research Across the Lifespan Program* to catalyse this field of research and develop badly needed expertise in low- and middle-income countries (LMICs). To mark this milestone, Fogarty felt it was important to review progress and consider how best to move forward.

In its first decade, the programme awarded about US\$84 million in more than 150 grants. Programme investigators have generated discoveries detailed in 435 peer-reviewed articles and 14 books or book chapters<sup>2</sup>. Scientists developed clinical assessment tools designed for low-resource settings, produced and tested novel interventions, and identified promising new approaches. The programme model has enabled investigators in the United States and other high-income countries to gain experience working in LMIC settings, while strengthening the research base of both US and LMIC institutions through collaborations.

**Given the scientific strides that have been made by working together, it is hoped that this research and training agenda can move forward in new and exciting directions.**

To help sustain this significant momentum, programme funding has supported long-term training of at least 138 LMIC scientists. Data generated by programme participants have provided crucial evidence that has been used to inform international and national practice and policy. Examples include identifying and helping to remove a global barrier to the availability of an anti-epileptic drug in Africa, increasing awareness of fetal alcohol syndrome on a national level in Russia, and convincing the Peruvian

government to institute acyclovir treatment for herpes simplex virus encephalitis. Finally, the initiative has built partnerships for research between US and foreign academic institutions, and strengthened the long-term research capacity of the LMIC institutions. It has extended the frontiers of neuroscience research to include many LMIC institutions and expanded the research workforce to more rapidly address some of the world's most pressing health problems.

The global brain programme would not have been possible without the full partnership of the other institutes at the NIH who joined forces to push forward the frontiers of brain research to include institutions and scientists in the developing world — the Eunice Kennedy Shriver National Institute of Child Health and Human Development, the National Eye Institute, the National Institute on Aging, the National Institute on Alcohol Abuse and Alcoholism, the National Institute on Deafness and Other Communication Disorders, the National Institute on Drug Abuse, the National Institute of Environmental Health Sciences, the National Institute of Mental Health, the National Institute of Neurological Disorders and Stroke, and the Office of Dietary Supplements.

In February 2014 funding partners, grantees and brain-disorder experts convened for several days of robust discussions to explore knowledge gaps and research opportunities. These conversations, organized by Fogarty's Center for Global Health Studies, were the genesis for this series of Reviews, which it is hoped will catalyse the international community to devote attention and resources to this crucial research agenda. Fogarty owes a debt of gratitude to all the participants, authors and the editors Donald Silberberg and Rajesh Kaloria.

Given the scientific strides that have been made in imaging, diagnostics, nanoscience, novel surgical interventions and genetics, by working together, it is hoped that this research and training agenda can move forward in new and exciting directions. Although the progress made is encouraging and mental health, substance misuse and chemical exposures are included in the United Nations proposal for sustainable development goals<sup>3</sup>, much more needs to be done.

It is hoped that this supplement will inspire other scientists and funding partners to join us in addressing the full spectrum of research, training, implementation and policy questions needed to alleviate the suffering caused by global brain disorders in LMICs and other countries around the world.

#### Roger I. Glass

Director, Fogarty International Center, National Institutes of Health, 31 Center Drive, Bethesda, Maryland 20892-2220, USA.

1. Whiteford, H. A. *et al.* Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010 *Lancet* **382**, 1575–1586 (2013).
2. Fogarty International Center. *Brain Disorders in the Developing World: Research Across the Lifespan Program Evaluation 2003–2013* <http://www.fic.nih.gov/About/Staff/Policy-Planning-Evaluation/Pages/fogarty-program-evaluation-brain-disorders.aspx> (Fogarty International Center, 2014).
3. United Nations. *Open Working Group proposal for Sustainable Development Goals* <https://sustainabledevelopment.un.org/> (UN, 2014).

#### COMPETING FINANCIAL INTERESTS

The author declares no competing financial interests. Financial support for publication has been provided by the Fogarty International Center.

#### ADDITIONAL INFORMATION



This work is licensed under the Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0>



INTRODUCTION **OPEN**

# Brain and other nervous system disorders across the lifespan — global challenges and opportunities

Donald Silberberg<sup>1</sup>, Nalini P. Anand<sup>2</sup>, Kathleen Michels<sup>3</sup> & Raj N. Kalaria<sup>4</sup>

This is an exciting time for scientific discovery that aims to reduce the frequency and impact of neurological, mental health and substance-use disorders. As it became increasingly clear that low- and middle-income countries have a disproportionate share of these disorders, and that many of the problems are best addressed by indigenous researchers who can seek context-sensitive solutions, the US National Institutes of Health and other research funders began to invest more in low- and middle-income country-focused research and research capacity-building to confront this significant public health challenge. In an effort to identify existing information, knowledge gaps, and emerging research and research capacity-building opportunities that are particularly relevant to low- and middle-income countries, in February 2014 the Center for Global Health Studies at the National Institutes of Health Fogarty International Center held a workshop to explore these issues with scientific experts from low- and middle-income countries and the United States. This evolved into the preparation of the Reviews in this supplement, which is designed to highlight opportunities and challenges associated with topical areas in brain-disorders research over the coming decade. This Introduction highlights some of the over-arching and intersecting priorities for addressing causes, prevention, treatment and rehabilitation as well as best practices to promote overall nervous system health. We review some brain disorders in low- and middle-income countries, while the Reviews describe relevant issues and the epidemiology of particular conditions in greater depth.

*Nature* 527, S151–S154 (19 November 2015), DOI: 10.1038/nature16028

This article has not been written or reviewed by *Nature* editors. *Nature* accepts no responsibility for the accuracy of the information provided.

The proportion of the global burden of disease (GBD) that is attributable to neurological, mental health, developmental and substance-use (NMDS) disorders is expected to rise worldwide, partly because of the projected increase in the number of individuals reaching the age at which they are at risk of onset of many of these disorders<sup>1,2</sup>. This rise will be steeper in low- and middle-income countries (LMICs) given the long-term effects of early life trauma, infectious disease and malnutrition, which contribute to the development of these disorders that in turn lead to early death or a lifetime of disability. Despite their significant contribution to the burden of disease and disability, NMDS disorders have been largely absent from the global health research agenda, and LMICs have insufficient capacity to address them. The past two decades have witnessed increased attention to, and investments in, NMDS issues. However, given the rising burden of NMDS disorders worldwide and opportunities to build on scientific advances while strengthening LMIC research capacity, we are at a crucial juncture in moving this agenda forward. This is a time to reflect on and use what we have learned to confront existing, and prepare for future, challenges. We hope that the research and research capacity-building priorities discussed in this collection will help to galvanize action to confront the rising tide of brain and other nervous system disorders.

## INTERNATIONAL HEALTH COMMUNITY RECOGNITION

Before the publication of the World Bank's 1993 seminal Annual

Report, *Investing in Health*, neurological, psychiatric, developmental and substance-use disorders in LMICs were considered unusual, difficult to understand and a low priority for research investment. The 1993 report, based largely on the first studies of the global burden of disease<sup>3</sup>, has since been refined and progressively updated. Prior to the implementation of measures such as the disability-adjusted life year (DALY; Table 1), the global burden of disease was primarily quantified in terms of mortality. With the advent of the DALY, the importance of neurological and psychiatric disorders became evident, accounting for approximately 28% of the global burden of disease. Subsequently, the US Institute of Medicine called for the need to direct research resources towards addressing NMDS disorders in LMICs<sup>4</sup> in its 2001 study and report, *Neurological, Psychiatric and Developmental Disorders — Meeting the Challenge in the Developing World*. This report not only encapsulated a growing body of evidence regarding the impact of brain disorders, but it also provided the seed for a new initiative to support research and research training in global brain disorders: the *Brain Disorders in the Developing World: Research Across the Lifespan Program* (brain programme) supported by the National Institutes of Health (NIH) Fogarty International Center, several NIH Institute and Center partners, and other organizations (Box 1). Further impetus was added by inclusion of several chapters on NMDS disorders in the 2006 second edition of *Disease Control Priorities in Developing Countries*<sup>5</sup>.

<sup>1</sup>Department of Neurology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. <sup>2</sup>Division of International Science Policy, Planning and Evaluation and Center for Global, Health Studies, Fogarty International Center, National Institutes of Health, Bethesda, Maryland 20892, USA. <sup>3</sup>Division of International Research and Training, Fogarty International Center, National Institutes of Health, Bethesda, Maryland 20892, USA. <sup>4</sup>Institute of Neuroscience, Newcastle University, Newcastle-upon-Tyne NE4 5PL, UK. Correspondence should be addressed to D. S. e-mail: silberbe@mail.med.upenn.edu or R. N. K. e-mail: r.n.kalaria@ncl.ac.uk.

**Table 1** | Proportions of age-standardized estimated disability adjusted life years (DALYs; >0.5%) attributable to all neurological, mental health and substance-use disorders by high-income and low- and middle-income countries (LMICs)

Disorder	Absolute DALYs* as global per thousand people (rank) †	Percentage in high-income countries (developed)	Percentage in low- and middle-income countries (developing)
<b>NEUROLOGICAL SEQUELAE IN VARIOUS INFECTIOUS, SYSTEMIC AND CONGENITAL DISORDERS</b>			
HIV/AIDS‡	130,900	1.19	3.67
All neglected tropical diseases and malaria‡	108,700	0.072	5.16
Nutritional deficiencies§	85,300 (3)	0.63	3.95
Low back pain	83,100 (1)	5.8	2.78
Neonatal encephalopathy (birth asphyxia/trauma)	50,150	0.33	2.33
Sensory organ diseases	34,700 (5)	1.68	1.35
Neck pain	33,640 (4)	2.24	1.14
Meningitis‡	29,400	0.13	1.38
Brain and nervous system cancers	6,100	0.53	0.19
Down's syndrome	1,775	0.8	1.71
<b>NEUROLOGICAL DISORDERS: NON-COMMUNICABLE¶</b>			
Cerebrovascular disease (all strokes)#	102,200	5.97	3.79
Neurological disorders (all)	73,800	4.42	2.71
Migraine	22,360 (6)	1.21	0.84
Epilepsy	17,400 (23)	0.44	0.75
Alzheimer's disease and other types of dementia	11,350 (21)	1.75	0.22
Other diagnosed neurological disorders	17,870	0.53	0.76
<b>MENTAL HEALTH AND BEHAVIOURAL DISORDERS</b>			
Mental health (all) **	185,200 (29)	11.1	6.73
Unipolar depressive disorders	74,260	4.01	2.81
Major depressive disorder	63,200 (2)	3.42	2.39
Anxiety disorders	26,830 (9)	1.6	0.98
Schizophrenia	15,000 (11)	0.85	0.49
Bipolar depressive disorders	12,870 (17)	0.63	0.5
Dysthymia	11,100 (16)	0.6	0.42
<b>SUBSTANCE-USE DISORDERS††</b>			
Drug-use disorders	20,000 (19)	1.53	0.67
Alcohol-use disorders	17,640 (22)	1.52	0.56
Total rate estimates (major categories with direct nervous-system involvement and >0.5% DALYs only in either a developed or developing region)	716,905	36.15	29.1

\*DALY is defined as a measure of overall disease burden and expressed as the sum of years of potential life lost due to ill-health, disability or premature mortality. DALYs<sup>1</sup>, listed high to low for each neurological, mental health and substance-use disorder type, for only those disorders with >0.5% rates in either high-income country or LMIC. Estimates for high-income countries were derived from the 'developed' region data whereas those for LMIC were from 'developing' region data<sup>1</sup>. †Ranked as disorders among the top 25 causes of global years lived with disability (YLDs)<sup>12</sup>. ‡Conduct disorder was foremost. §Includes cerebral malaria, encephalitis and HIV dementia. §Includes hearing and vision loss, including macular degeneration. ||Iron-deficiency anaemia is key. ¶Non-communicable neurological disorders include types of dementia, Parkinson's disease and related disorders, epilepsy, multiple sclerosis, migraine, tension-type headache and other neurological disorders. #Includes both ischaemic (70–80%) and haemorrhagic strokes (20–30%). \*\*Mental health includes major depressive disorder, dysthymia, unipolar and bipolar depressive disorders, schizophrenia, anxiety disorders, eating disorders, autism, Asperger's syndrome, attention deficit hyperactivity disorder, conduct disorder, idiopathic intellectual disability and other mental and behavioural disorders. ††Substance-use disorders include alcohol, opioid, cocaine, amphetamines, cannabis and other drug use.

Subsequently, The Grand Challenges Canada – Saving Brains and Global Mental Health, The EU Tropical Diseases Research Program, the UK Wellcome Trust, the UK Medical Research Council and others have joined to support global research and training to address brain disorders in LMICs (Fig. 1). Despite the successes and investments from these efforts and the willing support from numerous partners, including the US Agency for Overseas Development and the Bill and Melinda Gates Foundation, the implications of the burden of NMDs disorders in LMICs remains insufficiently addressed by appropriate policies and research and capacity-building investments. As we analyse these important ventures (Fig. 1), there are clearly several geographical gaps in investment, concerns for the extent of overlap in research funding for certain countries, and the potential for partnerships between funders and between countries.

## BURDEN AND RANGE OF NERVOUS-SYSTEM DISORDERS

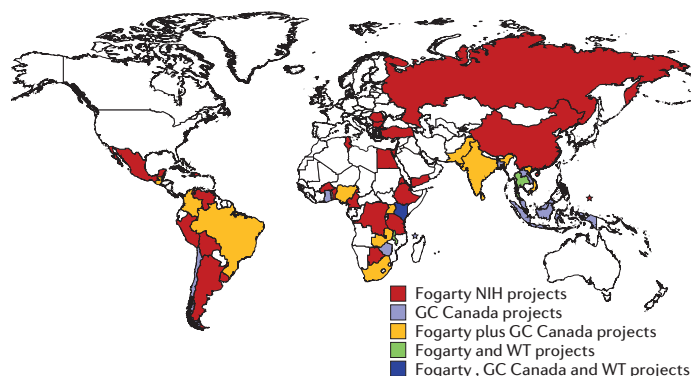
With the advent of the DALY, the importance of neurological and psychiatric disorders was undeniable; by 2010 all causes of NMDs disorders, including stroke, were estimated to account for more than 29% of the GBD (Table 1). In tandem with increased life expectancy<sup>6</sup>, the GBD has shifted from premature death to increased years lived in disability per 100,000. Not surprisingly, as the GBD has continued to move in the direction of non-communicable diseases, the burden of mental health and substance-use disorders has increased by around 40% in the past two decades<sup>7</sup>. Non-communicable diseases have been found to be the major cause of death and disability (Table 1), it is therefore time to include them as crucial priorities for research and policy initiatives in LMICs.

Although our appreciation of the importance of NMDs has matured significantly, serious contextual problems that obfuscate the true burden

## BOX 1 | THE GLOBAL BRAIN PROGRAMME

Scope of, and data from, 2003 to 2013 National Institutes of Health/Fogarty International Center ongoing Global Brain programme<sup>11</sup>.

- Support the development and conduct of innovative, collaborative research and research training projects, between developed and developing country scientists, on brain disorders throughout life, relevant to LMICs<sup>11</sup>
- The research should include a lifespan approach, for example nutrition or early exposure with sequelae or impact throughout life
- The programme has created a global network of researchers in 45 countries
- Projects have informed policies and programmes at the national and international levels
- Grants have resulted in 435 peer-reviewed publications from 249 unique journals
- The programme has catalysed new research projects supported by other funders
- Outputs include new tools for clinical assessment in the LMIC context, laboratory tools and methods



**Figure 1** | Research support for neurological, mental health and substance-use (NMDS) disorders in low- and middle-income countries. Colour-coded distribution of concerted programmes from principal agencies that support (currently and during the past decade) projects on NMDS disorders. Further details can be obtained from the agency reports<sup>11,13</sup>. There are numerous other organizations that support programmes that affect brain health, but that are not shown on this map. Fogarty, Fogarty International Center; GC Canada, Grand Challenges Canada — Saving Brains and Global Mental Health; WT, UK Wellcome Trust.

of these disorders remain. By definition, disorders in the GBD classification are attributed to underlying causes rather than to clinical manifestations. Thus, for example, disabilities that result from neuropathies are attributed to diabetes or HIV. Similarly, among types of epilepsy, only idiopathic epilepsies are considered to be neurological burdens. Where an underlying cause for seizures is known or suspected (for example, traumatic brain injury), the GBD methodology attributes the epilepsy-associated disability to the underlying cause. Highest-ranked causes of death in the GBD 2010 study include several that are neurological in nature, for example stroke (ranked 2nd), malaria (ranked 11th), neonatal encephalopathy (ranked 24th) and meningitis (ranked 29th). All these conditions, none of which are included in the neurological category, rank similarly high in terms of their contributions to global disability<sup>8</sup>.

Neglected tropical disorders (NTD) are also disproportionately neurological in nature<sup>9</sup>. Three of the seventeen NTDs recognized by the World Health Organization (WHO) are primarily neurological infections (rabies, human African trypanosomiasis and leprosy) and, of the remaining NTDs, some of the more severe manifestations are the result of nervous-system involvement (central nervous system schistosomiasis and Chagas-related stroke). When these are considered in this context, the true estimate of DALYs that are accounted for by NMDS disorders would be substantially higher than 29% (Table 1).

Research undertaken since the US Decade of the Brain campaign in the 1990s has established that many of the most common and disabling neurological conditions are preventable or remedial with inexpensive therapies. Treatment for epilepsy and secondary stroke prevention are ranked by the World Bank among the ‘best buys’ in global health<sup>10</sup>. In the past decade, relatively small investments in scientific inquiry that are relevant to brain disorders in LMICs are yielding crucial insights, which are applicable to the broader global community. These findings also point the way to the intervention studies that should follow, which range from definitive clinical trials to population-level interventions targeting risk factors for NMDS disorders. Where research has pointed toward more optimal ways to structure medical education or provide health services, implementation with rigorous evaluations is needed and plans for a broad scale-up delineated.

Given this unique point in time, in February 2014 the Center for Global Health Studies at the Fogarty International Center held a workshop to explore the state of the science and to identify emerging research and research capacity-building priorities in brain disorders that are particularly relevant to LMICs. This evolved into the preparation of the Reviews in this collection, which is designed to highlight opportunities and challenges associated with specific topical areas in brain-disorders

research in the coming decade, including causes, prevention, treatment and rehabilitation. We highlight some of the over-arching and intersecting priorities for addressing brain disorders in LMICs, and the subsequent Reviews describe these issues and the epidemiology of particular conditions in greater depth.

## THEMES AND INTERSECTING RESEARCH APPROACHES

Despite financial and human-resource challenges, there have been many exciting discoveries related to NMDS disorders that have global implications. Examples include refinement of the classification of HIV/AIDS; realization of more substance-use disorders; definition of the sequelae of infectious diseases, seizure disorders and behavioural disorders; and identification of mechanisms of environmental toxicants in sub-Saharan Africa. Similarly, genetic studies associated with neurodegenerative disorders, for example Huntington’s disease and Alzheimer’s disease in Venezuela and Columbia have advanced our knowledge of diagnosis, risk factors and prospective treatments. Genetic and cognitive studies in Brazil are leading to interventions that may help to mitigate cognitive deficits in children owing to malnutrition.

The unique genetic makeup, environmental and traumatic exposures, infections, the local health system and nutritional challenges of diverse LMIC settings, mean that interventions and approaches that have been developed and tested in high-income country settings will often not be feasible or effective if simply transplanted to LMICs. In fact, solutions developed in LMICs may have global applications, such as new surgical techniques for children with incipient hydrocephalus developed in Uganda (see page S155) and may lead to a decrease or an elimination of disability caused by this condition, and perhaps cheaper and better treatments for children around the world. Thus, it is crucial that research be conducted in the environments in which the techniques are intended to be used. Several Reviews in this series provide examples of why this is the case in the context of brain disorders (for example, disorders associated with development, trauma, environment and adolescence). The conduct of high-quality research requires a cadre of in-country and well-trained scientists, and an adequately resourced research infrastructure that can support their research (see page S207). This series highlights priorities for research and research training that are related to specific diseases and conditions. We highlight overarching priorities included in these papers that cut across multiple disease areas (Box 2).



## BOX 2 | PRIORITIES FOR DIFFERENT THEMES ACROSS MULTIPLE DISORDERS

### Surveillance and epidemiology

- Determination of prevalence and incidence of specific brain disorders, including in vulnerable populations
- Strengthening of reporting infrastructure, data quality and disease registries, and standardization of registries for country comparisons
- Use of statistical modelling where possible to inform resource allocation and implementation of interventions
- Prevalence of specific risk factors and measures of their impact among affected populations
- Better design of disease epidemiology that accurately captures incidence, type and duration of nervous system sequelae of infections and other conditions

### Basic and clinical research

- Understanding the epigenetic effects on the nervous system resulting from malnutrition, infections, environmental exposures and psychosocial factors
- Better understanding of the influence of genetics and the genome on brain-disorder pathogenesis and progression
- Explore the relationship between an individual's and a population's microbiome and correlates with the presence of diverse disorders
- Understanding of pathways that lead to late onset of neurological disease owing to early exposures
- Advance understanding of gene–environment–brain interactions
- Assessment of risks and interactions of co-infections and co-morbidities
- Development of effective and feasible physical, occupational and cognitive rehabilitation interventions
- Creation of biobanking, including of brain tissues in LMICs.

### Implementation science and health systems

- Engagement of diverse stakeholders (including decision makers and programme implementers) to facilitate uptake and scale-up of effective interventions
- Evaluation of task-sharing models to address health-care provider shortages and to help create efficiencies in the health-care system and facilitate the delivery of effective services
- Research on how best to scale-up interventions that have proven efficacious in smaller, controlled settings

### Technology advancement

- Development and refinement of effective point-of-care diagnostic tools, particularly for genetic-based neurological disorders
- Development and testing of mobile-technology interventions to screen, diagnose and monitor treatment of brain disorders
- Development and implementation of low-cost tools and techniques for the treatment and rehabilitation of conditions arising from brain and other nervous-system trauma

### Capacity-building and research-infrastructure needs

- More clinician–neuroscientist researchers and others who are trained to deliver appropriate care for patients with brain disorders and participate in conducting research
- Strengthen capacity in neuroethics
- Development of enhanced and culturally adaptable cognitive assessment and screening tools
- Increased laboratory capacity
- Increased access to electrodiagnostics, genomic sequencing and neuroimaging technologies

## WHAT NEXT?

Resource constraints and the unique factors that people living in LMICs face impose a particularly onerous burden on these countries, where most of those with brain and other nervous-system disorders live. This burden significantly affects the ability of children and adolescents to thrive and live out their true potential, and the ability of young adults to be economically productive and support their families, as well as the opportunity for older adults to age in safe and nurturing settings. However, we may be at a tipping point for research related to global brain disorders. Over the past few decades, exciting basic science discoveries have been made, effective interventions have been developed and advances in technology have set the stage for a research agenda that can lead to unprecedented progress in this field. In addition, research capacity strengthening in LMICs and an opportunity for comparative studies of the markedly different environments that exist between high-income countries and LMICs will yield universally beneficial knowledge.

Reduction of disease and disability that are associated with brain and other nervous-system disorders over the next decade will demand increased engagement from, and collaboration among, the scientific community, research funding agencies, national governments, academic institutions, multilateral organizations, advocacy organizations and health providers to increase both research in LMICs and the indigenous research capacity. We need to build on current knowledge of overall neurological health, and improve the lives of those living with brain and other nervous system disorders at any stage of life.

1. Murray, C. J. *et al.* Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* **380**, 2197–2223 (2012).
2. Spencer, S. Global Burden of Disease 2010 Study: a personal reflection. *Glob. Cardiol. Sci. Pract.* **2013**, 115–126 (2013).
3. Murray, C. J., Lopez, A. D. & Jamison, D. T. The global burden of disease in 1990: summary results, sensitivity analysis and future directions. *Bull. World Health Organ.* **72**, 495–509 (1994).

4. Institute of Medicine. *Neurological, and Psychiatric and Developmental Disorders Disorders, Meeting the Challenge in the Developing World* (National Academy, 2001).
5. Jamison, D. T. *et al.* *Disease Control Priorities in Developing Countries* (Oxford Univ. Press and The World Bank, 2006).
6. Salomon, J. A. *et al.* Healthy life expectancy for 187 countries, 1990–2010: a systematic analysis for the Global Burden Disease Study 2010. *Lancet* **380**, 2144–2162 (2012).
7. Whiteford, H. A. *et al.* Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *Lancet* **382**, 1575–1586 (2013).
8. Chin, J. H. & Vora, N. The global burden of neurologic diseases. *Neurology* **83**, 349–351 (2014).
9. World Health Organization. *The 17 neglected tropical diseases* [http://www.who.int/neglected\\_diseases/diseases/en/](http://www.who.int/neglected_diseases/diseases/en/) (WHO, 2014).
10. World Economic Forum and World Health Organization. From burden to 'best buys': reducing the economic impact of non-communicable diseases in low- and middle-income Countries [http://www.who.int/nmh/publications/best\\_buys\\_summary/en/](http://www.who.int/nmh/publications/best_buys_summary/en/) (WHO/WEF, 2011).
11. Fogarty International Center. *The Global Brain Program* [www.fic.nih.gov](http://www.fic.nih.gov) (FIC, 2015).
12. Global Burden of Disease Study. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* **386**, 743–800 (2015).
13. Grand Challenges Canada. *Grand Challenges Canada — Saving Brains* <http://www.grandchallenges.ca/saving-brains/> (2015).

### ACKNOWLEDGMENTS

We would like to thank C. Wolfman, for her exceptional coordination, research and editorial support that helped to make this supplement possible. We are also grateful to several advisors and specially commissioned reviewers of the manuscripts for their invaluable constructive comments and suggestions in improving the articles.

### COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests. Financial support for publication has been provided by the Fogarty International Center.

### ADDITIONAL INFORMATION



This work is licensed under the Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0>

## REVIEW OPEN

# Reducing neurodevelopmental disorders and disability through research and interventions

Michael J. Boivin<sup>1,2</sup>, Angelina M. Kakooza<sup>3</sup>, Benjamin C. Warf<sup>4</sup>, Leslie L. Davidson<sup>5,6</sup> & Elena L. Grigorenko<sup>7</sup>

We define neurodevelopment as the dynamic inter-relationship between genetic, brain, cognitive, emotional and behavioural processes across the developmental lifespan. Significant and persistent disruption to this dynamic process through environmental and genetic risk can lead to neurodevelopmental disorders and disability. Research designed to ameliorate neurodevelopmental disorders in low- and middle-income countries, as well as globally, will benefit enormously from the ongoing advances in understanding their genetic and epigenetic causes, as modified by environment and culture. We provide examples of advances in the prevention and treatment of, and the rehabilitation of those with, neurodevelopment disorders in low- and middle-income countries, along with opportunities for further strategic research initiatives. Our examples are not the only possibilities for strategic research, but they illustrate problems that, when solved, could have a considerable impact in low-resource settings. In each instance, research in low- and middle-income countries led to innovations in identification, surveillance and treatment of a neurodevelopmental disorder. These innovations have also been integrated with genotypic mapping of neurodevelopmental disorders, forming important preventative and rehabilitative interventions with the potential for high impact. These advances will ultimately allow us to understand how epigenetic influences shape neurodevelopmental risk and resilience over time and across populations. Clearly, the most strategic areas of research opportunity involve cross-disciplinary integration at the intersection between the environment, brain or behaviour neurodevelopment, and genetic and epigenetic science. At these junctions a robust integrative cross-disciplinary scientific approach is catalysing the creation of technologies and interventions for old problems. Such approaches will enable us to achieve and sustain the United Nations moral and legal mandate for child health and full development as a basic global human right.

*Nature* 527, S155–S160 (19 November 2015), DOI: 10.1038/nature16029

This article has not been written or reviewed by *Nature* editors. *Nature* accepts no responsibility for the accuracy of the information provided.

One evaluation of early childhood developmental status in low- and middle-income countries (LMICs) estimates that 15.7% of children are significantly delayed in their cognitive development, 26.3% in socioemotional development and 36.8% in either or both (D. C. McCoy, personal communication). Stunting, low wealth and living in a rural area are significantly associated with neurodevelopmental delay; most of the children live in Africa and eastern Asia. Fortunately, neurodevelopmental science is benefitting from rapidly expanding technologies for the integration of the environmental (for example, infectious disease, nutritional and carer quality), brain-related (for example, developmental neuroscience and brain imaging) and genetic (for example, epigenetic modelling and genomic big data) domains that drive neurodevelopment. Figure 1 illustrates the mutually interactive nature of these three developmental domains, along with the current strategic areas of research at the environment–brain–gene interface (Box 1).

Advances in developmental science have triggered a reconceptualization of neurodevelopment based on the recognition that

developmental processes are a part of child health in the broader context of communicable and non-communicable disease<sup>1</sup>. The developmental origins of the health and disease hypothesis proposes that the physiological processes of developmental plasticity operate in early childhood, but have the potential for adverse consequences in later life<sup>2</sup>. Consequently, childhood – particularly early childhood – is a high-priority target for both preventive and remediating interventions to address the pervasive developmental needs in LMICs (D. C. McCoy, personal communication).

In this Review, we describe several high-impact findings that have emerged from research in low-resource settings that pertain to the developmental milieu of the child, its relationship to the brain and behavioural neurodevelopmental integrity of the child (neurodevelopmental disorders), and the genetic and epigenetic underpinnings that can drive this relationship. As we review key scientific advances in each of these three domains, we propose strategic areas of ongoing and future research that could provide innovative models to fuel significant

<sup>1</sup>Department of Psychiatry, College of Osteopathic Medicine, 965 Fee Road, East Fee Hall Room 225, Michigan State University, East Lansing, Michigan 48824, USA.

<sup>2</sup>International Neurologic and Psychiatric Epidemiology Program (INPEP), West Fee Hall Room 321, Michigan State University, East Lansing, Michigan 48824-1315, USA.

<sup>3</sup>Department of Paediatrics and Child Health, School of Medicine, Makerere University College of Health Sciences, PO Box 7072, Kampala, Uganda. <sup>4</sup>Boston Children's Hospital, Department of Neurosurgery, 300 Longwood Avenue, Hunnewell, 2nd Floor, Boston, Massachusetts 02115, USA. <sup>5</sup>Department of Epidemiology, Mailman School of Public Health, College of Physicians and Surgeons at the Columbia University Medical Center, 722 West 168 Street 1613, New York, New York 10032, USA. <sup>6</sup>Department of Pediatrics, College of Physicians and Surgeons at the Columbia University Medical Center, 722 West 168 Street 1613, New York, New York 10032, USA. <sup>7</sup>Yale Child Study Center, Department of Psychology, Department of Epidemiology and Public Health, Yale University, PO Box 207900, 230 South Frontage Road, New Haven, Connecticut 06520-7900, USA. Correspondence should be addressed to M. J. B. e-mail: boivin@msu.edu.

advances and evidence-based interventions for meeting the developmental needs of children. We conclude by summarizing ways in which this model (environment, brain and gene) provides rich opportunities for a more global approach to child-development science, making it possible to achieve the UNICEF mandate of full child health and development for all<sup>3,4</sup>.

## APPROACHES TO MALARIA

In 2013 there were around 198 million cases of malaria of which 584,000 resulted in death<sup>5</sup>. A child dies from malaria every minute, and one in four survivors present with significant neurodevelopmental impairment<sup>6,7</sup>. However, cognitive rehabilitation, speech and physical therapy, and carer-training interventions can improve cognitive performance and behaviour of treated mother-child pairs<sup>8,9</sup>. As rehabilitation approaches are evaluated, there is mounting evidence of the neurocognitive benefits of computerized cognitive rehabilitation training (CCRT) in African children with a brain injury as a result of severe malaria and in those with HIV-related brain injury<sup>10</sup>. Dissemination and implementation science must now inform innovative approaches to bring such interventions to scale in low-resource communities. Mobile network health (mHealth) research opportunities are a high priority, given the ever-increasing access that children and adolescents in low-resource settings have to mobile-based internet and computing technologies<sup>11</sup>. Another key strategic research opportunity is to evaluate the impact of neurocognitive rehabilitation interventions such as CCRT, on the enhancement of brain-development neuroprotective factors<sup>12,13</sup> (Box 2). We can then evaluate the extent to which such brain-based biomarkers mediate the neuropsychological benefits of CCRT, along with how neurocognitive rehabilitative interventions diminish biomarkers of brain inflammation (such as tumour necrosis factor- $\alpha$  (TNF- $\alpha$ ) and creatinine).

## APPROACHES TO PAEDIATRIC HIV

Globally, roughly 3.4 million children live with HIV infection and are at high risk of significant neurodevelopmental disabilities. Of these, almost 90% live in Africa where only 24% of infected children have access to anti-retroviral (ARV) treatment<sup>14</sup>. These children's environmental risk factors are compounded by poor nutrition owing to protein and specific micronutrient deficiencies<sup>15</sup>. They also often have parasitic, respiratory and enteric diarrheal infections<sup>16</sup>. Such compounded risk exists whether a child is infected with HIV (proximal risk) or lives in a household or community where HIV has a persisting and significant disruptive impact (distal risk)<sup>17</sup>. Multifaceted risk for all kinds of early developmental insults (for example, infection, malnutrition and poverty) demands that children in low-resource communities need a comprehensive package of assessments and interventions to holistically enhance their development<sup>16</sup>.

Recent epigenetic evidence suggests that chronic poverty may 'shrink' children's brains over successive generations as documented by longitudinal multigeneration brain-imaging research<sup>18</sup>. These considerations justify the junction between the environment and the brain as a highly strategic point of intervention. This is further illustrated by the strategic importance of implementation-science research in designing an effective comprehensive package of services for antenatal and postnatal care. This is evident when considering at-risk adolescent mothers in LMICs and the heightened risk of neurodisability in their infants. Pregnancy in adolescence is associated with premature delivery, stillbirth, fetal distress, birth asphyxia, low birth weight and miscarriage<sup>19</sup>. Furthermore, if the adolescent mother is also suffering from malnutrition these risks are compounded for the infant — long-term effects as a consequence of low birth weight include stunting, poor neurodevelopmental outcomes, and increased susceptibility to cardiovascular and metabolic diseases such as obesity and diabetes<sup>20</sup>. In fact, there is evidence that environmental factors such as nutrition can alter epigenetic modifications and thus play a part in the development of these disorders later in life<sup>21</sup>. The maternal

microbiome is also important to infant health outcomes, including the risk of pre-term birth, the development of gastrointestinal diseases such as irritable bowel syndrome, and the development of the immune system<sup>22</sup>.

One of the greatest public health initiatives developed in the modern era of infectious disease is the prevention of mother-to-child transmission (PMTCT) of HIV. These interventions have reduced perinatal infection of children born to infected mothers from more than 30% to less than 1%<sup>23</sup>. However, there are still gestational neurodevelopmental risks associated with early exposure to ARVs<sup>24</sup>. One of the most exciting developments in the treatment of HIV has been the development and anti-retroviral characterization of VRC01. This is a potent and broadly neutralizing anti-HIV monoclonal antibody that prevents HIV-1 transmission from plasmacytoid dendritic cells to CD4 T lymphocytes<sup>25</sup>. Once proven safe for infants, such therapies should be administered as soon as possible after the diagnosis of HIV in infants, and the long-term neurodevelopmental and neurocognitive protective benefits of such innovative treatment strategies should be evaluated. These therapies could also be effective in the prevention of HIV transmission.

## TRAUMA-ASSOCIATED PSYCHIATRIC ILLNESS

Another strategic research opportunity is to further evaluate how maternal depression is associated with widespread changes in DNA methylation in their offspring<sup>26,27</sup>. Such epigenetic processes can result in heightened risk of depression and anxiety disorders in children as they become adults<sup>28</sup>. How best to package and bring to scale a strategic set of intervention services that address this remains a neglected area in high-impact implementation science. Likewise, populations traumatized through conflict and genocide can pass on psychiatric disorders transgenerationally. This may be partly mediated by the hormonal effects of maternal stress on neuropsychiatric risk for children *in utero* in regions where women have been traumatized through sexual violence in conflict zones (glucocorticoid-mediated induction of cytokine inflammatory responses causing methylation of DNA in children *in utero*). Such intergenerational epigenetic mechanisms of psychiatric disorders necessitate evidence-based and sustainable community-wide treatment strategies to address these disorders within the foundational mother-child caring fabric of that society<sup>29</sup>. Task shifting will be a crucial strategy in addressing such community mental health support services<sup>30,31</sup>. There is evidence to support the effectiveness of a year-long maternal carer training programme for children who are affected by HIV in rural Uganda to facilitate child development in low-resource settings, while remediating maternal depression and enhancing carer functionality<sup>8,11</sup>.

## NODDING SYNDROME

The beginning of the millennium was marked by the manifestation of the enigmatic condition nodding syndrome, which affects school-age children, and is reported in South Sudan, northern Uganda and southern Tanzania. This condition is characterized by episodes of repetitive nodding (dropping forward of the head) often coupled with seizure-like behaviours (for example, convulsions or staring spells) that occur during attempted feeding<sup>32,33</sup>. Nodding syndrome is also characterized by stunted brain growth, including significant brain atrophy near the hippocampal and glia matter of the brain and significant cerebellar involvement. This is accompanied by lifelong profound neurodisability, severe behavioural problems and high mortality<sup>34</sup>.

The nodding is caused by an atonic seizure, but the aetiology of this seizure is unknown, although associations with other developmental conditions have been established. Nodding syndrome is most prevalent in areas with high infection rates of the parasitic worm *Onchocerca volvulus* — a nematode carried by black fly of the genus *Simulium* — the bites of which can cause onchocerciasis, a highly prevalent type of blindness caused by infection. Other reports suggest an association between the syndrome and malnutrition<sup>35</sup>. Future research of this



syndrome must focus on understanding the aetiology so that it can be prevented, diagnosed early and treated effectively. Emerging diseases that profoundly affect children, such as nodding syndrome, provide an important opportunity for developing diagnostic, management and intervention techniques adapted to LMICs that, in turn, can be used for the prevention of worldwide outbreaks of diseases that lead to severe disability.

Although nodding disease is highly localized, such enigmatic disorders that arise from time-to-time and result in profound neurodisability are important because they reveal the urgent need to develop scientific models that can be seamlessly integrated into emerging disciplines. These include geographical ecological mapping, maps of parasitology dispersion, genotypic mapping across populations and their geographical dispersions, and geographically mapped epidemiological risk of infectious disease. These multi-layered models must then be integrated with neuropathogenic mechanism models that include sensitive and specific brain inflammatory markers, as well as the corresponding neuropsychological sequelae of such central nervous system inflammatory markers.

## MALNUTRITION AND DISEASE

Childhood malnutrition, both through prenatal and perinatal maternal micronutrient deficiencies<sup>36</sup>, infant micronutrient deficiencies<sup>37</sup>, and protein-calorie deficiency, imposes a heavy burden on neurodevelopment<sup>38–40</sup>. The primary effects of malnutrition have been associated with elevated mortality, morbidity, and risk of cognitive and socio-emotional impairment. Although it has been extensively researched, and interventions have been attempted, malnutrition remains a serious challenge to children's development in LMICs. Efforts have not yet succeeded in eliminating malnutrition or in successfully bringing interventions to scale<sup>41</sup>. Secondary effects of malnutrition are associated with vulnerability to microbial pathogens that can also severely disrupt neurodevelopment<sup>42,43</sup>.

### Enteric infections

The aetiology of malnutrition is complex. In particular, malnutrition might result from enteric infections of bacteria that are highly prevalent in LMICs, and include both well-known (*Escherichia coli*, *Vibrio cholerae*, and species of *Salmonella*, *Shigella* and anaerobic streptococci)<sup>44</sup> and emerging pathogens (enteroaggregative *E. coli*, *Cryptosporidium* and *Giardia*)<sup>45</sup>. These infections can significantly affect childhood brain or behavioural development, presumably through damage to the gut microbiota. This can lead to intestinal inflammation that diminishes intestinal absorption, and protein and micronutrient deficiencies compounded by recurring dehydration and malaise<sup>46,47</sup>. This field of research has also significantly advanced our understanding of the inter-relationships between genetics (for example, neuroprotective *APOE* polymorphisms), enteric diseases, nutritional malabsorption and neurodevelopment in young children<sup>48,49</sup>.

An important research opportunity provided by this work involves the clinical evaluation of the neurodevelopmental benefits of micronutrient interventions to enteric disease, including whether glutamine works better than glucose as a key ingredient of oral rehydration and repair therapy (ORRT)<sup>50</sup>. Glutamate intervention may be more effective in the repair of intestinal barrier functions and hence improve child development as well as the absorption of ARV drugs in children with HIV.

### Food-borne neurotoxins and nutritional malabsorption

Konzo disease is a permanent, irreversible, upper-motor neuron disorder, occurring primarily in rural areas of sub-Saharan Africa that are dependent on bitter varieties of cassava (*Manihot esculenta*; an annual crop cultivated for its edible starchy tuberous root, which is a major source of carbohydrates and, therefore, a food staple). Epidemiological studies have documented konzo outbreaks — mostly in women and children — in periods of food insecurity that have been brought about by drought, displacement by war or conflict, or other factors that have

## BOX 1 | STRATEGIC ONGOING AND IMMINENT RESEARCH OPPORTUNITIES

Strategic ongoing and imminent research opportunities at the intersection between brain, gene and environment (Fig. 1), which potentially lead to neurodisability interventions in low- and middle-income countries.

### Gene and brain — ongoing research opportunities

- Neural tube defects with associated hydrocephalus and developmental brain anomalies.
- Monogenic disorders (either heritable or *de novo*) for which neurodevelopmental disorders are caused by various types of mutations (from a point mutation such as in sickle cell disease to in a single gene such as in Rett syndrome).
- Disorders due to alterations in the mitochondrial genome (for example, creatine deficiency syndromes).
- Neurodevelopmental disability of unknown origin such as autism spectrum disorders, specific learning disabilities, attention deficit hyperactivity disorder and conduct disorders.

### Environment and brain — ongoing and imminent (shown by \*) research opportunities

- Brain injury from central nervous system infections related to neonatal sepsis (meningitis, ventriculitis and cerebritis) and resulting post-infectious hydrocephalus.
- Toxic exposure of cyanide to young children fed cassava as a result of food insecurity and insufficient processing of cassava with high linamarin content.
- \*Evaluate neurocognitive rehabilitation interventions on brain development neuroprotective factors and on biomarkers of brain inflammation (for example, TNF- $\alpha$  and creatinine).
- \*Hormonal effects of maternal stress on child neuropsychiatric risk *in utero* for mothers in LMICs who have been traumatized through sexual violence in conflict zones (glucocorticoid-mediated inducer of the cytokine inflammatory responses).
- \*Chronic poverty may 'shrink' brains over successive generations as documented by magnetic resonance imaging research.

### Environment and gene — imminent research opportunities

- Maternal depression is associated with widespread changes in DNA methylation in their offspring that may persist into adulthood for exposed children.
- Genetic vulnerability to onchocerciasis may lead to neuroinflammation, seizures, and profound neurodisability in the form of nodding disease in select vulnerable populations in regions highly endemic for onchocerciasis.
- Genetic factors related to why asymptomatic positive malaria parasitaemia progresses to cerebral malaria (or severe malaria anaemia) in African children with subsequent brain injury.
- *APOE* and neuroprotection for enteric diseases (with elevated risk for age-associated dementia types).

led to the insufficient processing of cassava tubers. The insufficient breakdown of linamarin compounds that contain cyanide result in neurological damage and seem to lead to outbreaks of konzo, which has been documented mostly in the Congo, Central African Republic, Mozambique and Tanzania<sup>51–53</sup> with a prevalence of between 0.1% and 17% in affected villages<sup>54</sup>. Studies have recently documented neurocognitive impairments in children with konzo. Furthermore, even children who do not show signs of konzo, but who live in konzo-affected households may have neurocognitive impairment of working memory and learning ability<sup>55</sup>.

Konzo offers an important opportunity for integrative neurodevelopmental science. Neuroinflammatory markers of brain injury from

cyanide toxicity and inflammatory markers of microbiota destruction in the gut from cyanide toxicity need to be mapped on sensitive neurocognitive impairment indicators in children. Konzo offers a rare opportunity to test integrative models of nutritional toxicity in the brain and gut against a backdrop of malnutrition and corresponding micronutrient deficiencies. Gauging their comparative weighting in the mediation of neurodevelopmental disability within the cognitive and neuromotor domains will allow us to determine the effectiveness of prevention and treatment strategies. Since konzo is entirely preventable, health education and promotion intervention methods should be evaluated at the community-wide level in terms of the benefit to disability-adjusted life years<sup>56</sup>.

## TREATING HYDROCEPHALUS IN LMICS

Hydrocephalus, the abnormal accumulation of cerebrospinal fluid in the cerebral ventricles, has multiple causes, and is especially prevalent in LMICs. Failure to treat the condition almost always leads to death or severe neurodevelopmental disability. Higher birth rates and limited perinatal care contribute to a greater burden of care for hydrocephalus in LMICs<sup>57</sup> (for example, there are 100,000–250,000 new infant cases of hydrocephalus annually in sub-Saharan Africa alone<sup>41</sup>). In addition to the expected burden of congenital hydrocephalus in LMICs, climate-driven neonatal ventriculitis of unknown pathogenesis has recently been identified as one of the chief causes of infant hydrocephalus (60% of cases in Uganda)<sup>58–61</sup>. In sub-Saharan Africa, rates of neonatal sepsis are estimated to be 170 per 1,000 births, with a corresponding mortality of 10 deaths per 1,000 births<sup>62</sup>. For survivors with post-infectious hydrocephalus (PIH), neurodevelopmental consequences of the primary brain injury can be devastating even before hydrocephalus develops. One-third of those with PIH remain profoundly disabled at five years, even after successful surgery<sup>63</sup>. However, innovative surgical techniques have been pioneered and developed in Uganda that have revolutionized the treatment of hydrocephalus worldwide<sup>63–65</sup>.

The standard treatment for hydrocephalus has long been the implantation of tubing that drains cerebrospinal fluid from the ventricles to the peritoneal cavity (ventriculoperitoneal shunt). However, this treatment creates lifelong dependence on an unreliable implanted device that often requires an emergency operation when it fails (40% failure within 2 years of the original implantation)<sup>66</sup>. An effective, minimally invasive treatment method (endoscopic third ventriculostomy (ETV) combined with endoscopic choroid plexus cauterization (CPC)) that avoids shunt-dependence in most infants was developed in Uganda as an alternative<sup>65</sup>. The procedure — the safety and efficacy of which were demonstrated initially in LMICs and then in high-income countries<sup>65</sup> — combines two techniques. These involve creating a new opening through the floor of the third ventricle and reducing the choroid plexus tissue in the lateral ventricles by cauterization. Building the capacity to achieve universal access to optimal and affordable hydrocephalus treatment for infants in LMICs is an ongoing challenge<sup>64</sup>. Present efforts involve task shifting by training non-physician medical officers to undertake the shunt placement, allowing neurosurgeons to focus on the more complex third ventriculostomy procedures. Dissemination and implementation research is needed to test the effectiveness of this task-shifting approach.

## GENETIC STUDIES OF NEURODEVELOPMENT

The genome has a substantial role in the aetiology of neurodevelopmental disorders. These disorders can be classified into six major categories (Box 2). There is abundant evidence that the disorders in all six categories may affect many facets of child development. The disorders in categories 1–3 are typically severe and impose multiple developmental challenges from birth. These conditions are referred to as congenital conditions (their broad definition also includes conditions that result from various challenges in pregnancy, such as severe micronutrient deficiency, for example folate deficiency). Given what is known about the prevalence of these conditions in high-income countries,

### BOX 2 | SIX GENETIC CAUSES OF NEURODISABILITY IN CHILDREN

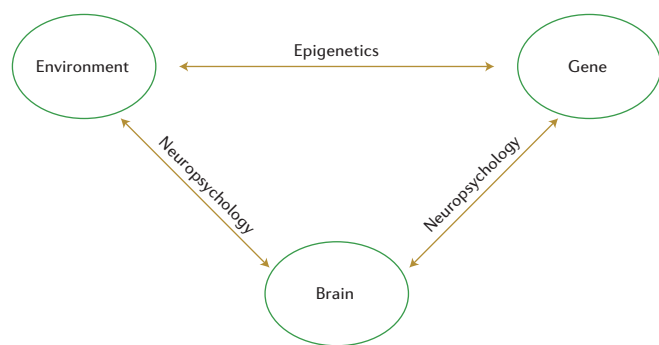
1. Disorders caused by specific genomic lesions (either heritable or arising *de novo*).
2. Monogenic disorders (either heritable or *de novo*) for which the disorders are caused by various types of mutations (from a point mutation, such as sickle cell disease, to repeat expansion) in a single gene (for example, mutations in the *MECP2* gene in Rett syndrome and a number of CGG repeats in the *FMR1* gene in fragile X syndrome).
3. Disorders due to alterations in the mitochondrial genome (for example, creatine deficiency syndromes).
4. Relatively common disorders such as autism spectrum disorders, specific learning disabilities, attention-deficit hyperactivity disorder and conduct disorders.
5. Disorders triggered by the environment, but the burden of which is controlled by the genome.
6. Conditions that arise from the involvement of the epigenome (modifications in the function of the genome that are not caused by any structural alterations in the genome itself).

estimates suggest that at least 7.6 million children are born annually with severe congenital conditions, and that the number is especially high in LMICs<sup>67</sup>.

The disorders in categories 4–6 include common multifactorial conditions with onset in early childhood. Of note, less than 50% of countries have policies for the control of these conditions<sup>67</sup>. These conditions currently constitute a substantial health challenge in high-income countries, but are substantially understudied, under diagnosed and underserved in LMICs. Although limited, the relevant research in LMICs unfolds in a number of dimensions, converging around the understanding that economic development and changes in lifestyle have led, or are leading to, a rapid increase in the observed prevalence of these multifactorial disorders. In other words, as people's environment improves, the role and prominence of genetic and genomic factors will increase. Common disorders include conditions that are attributable to epigenetic influences<sup>68</sup> (for example, DNA methylation and histone modification).

In this context, two epigenetic mechanisms have been highlighted. The first mechanism connects nutritional challenges to the manifestation of metabolic syndromes. This happens through a causal link between nutrient restrictions *in utero* and in early childhood, lack of clean water and sanitation, and high levels of infectious organisms in the environment. These can lead to epigenetic changes in pathways related to metabolism, blood pressure and glucose regulation<sup>69</sup>. The second mechanism is the link between psychological stress and the glucocorticoid-mediated inducer of the cytokine inflammatory response<sup>70</sup>. Both exemplify the developmental origins of the health-and-disease hypothesis and its relevance to the aetiology of neurodevelopmental disability in LMICs<sup>71</sup>.

Key research and training priorities related to these six disorder categories are: determining their global prevalence; training scientists in appropriate molecular technologies and sustaining this increased human-resource capacity by providing ongoing support and training to keep up with the rapid technological advances in the field; developing methods for cheap and reliable diagnoses of the widest possible range of congenital conditions and identification of the broadest possible range of risk factors for complex multifactorial disorders; developing practical, accessible and inexpensive procedures for family-planning counselling (preconception and post-delivery); and continuing to build the capacity and infrastructure needed to initiate cutting-edge, relevant research that is comparable with that taking place in high-income countries. These research and training priorities should translate



**Figure 1** | Mutually interactive domains — environment, gene and brain — interact in terms of environment–gene socioevolutionary processes (epigenetic), environment–brain moment-by-moment neurocognition (neuropsychology), and gene–brain universal brain and behavioural processes in child neurodevelopment. The foundation for this multi-level interaction is brain plasticity as shaped by risk and resilience in child neurodevelopment, which occurs at the evolutionary (physical environment), cultural (social environment), individual (brain) and neuronal genotype (genetic) levels. The most strategic points of research opportunity as presented in this Review occur at the intersections between brain, gene and environment.

into public health services that can help couples in family planning and the resource mobilization needed to nurture children with such disorders.

## CONCLUSION

We have outlined significant scientific findings and challenges that have emerged from research in LMICs. We have provided strategic research examples and areas of research opportunity (aetiology and intervention) at the junction between the environment, brain and gene. The dynamic interactions among these three domains are at the foundation of brain neurodevelopment in children (Box 1). New technologies are providing ever more sensitive biomarkers that can be related to the brain and behavioural neurodevelopmental integrity of the child. New technologies are also emerging that link the regional and global surveillance of neurodisability to environmental risk, and these can be integrated with the genetic and epigenetic underpinnings that drive this relationship.

Future approaches must accommodate the use of new data gathered by innovative technologies, offering fresh approaches to old problems in child development in LMICs. These new approaches will prove to be especially strategic at the points of interface and integration between the environment, gene and brain (Fig. 1). New models that can effectively integrate these three domains into a comprehensive and cohesive paradigm must have the following hallmarks.

## Interdisciplinary approaches

Research in LMICs needs to take into account the complex multifactorial causes of neurodisability (Fig. 1). Environmental risk from natural disasters, social unrest, poverty and infection can offset a child's neurodevelopmental trajectory at the points of interface between gene and brain, during the antenatal and postnatal stages of development. In addition to toxins from the diet such as cassava-based cyanide in konzo, environmental toxins from mining and biomass fuels are leading to levels of exposure that can affect child neurodevelopment across entire communities. These complex systems of developmental risk factors call for comprehensive interdisciplinary approaches to understand the developmental trajectories so as to identify children at serious risk of neurodisability and those in need of intervention in LMICs. For example, maternal health programmes need to work closely with early childhood programmes to ensure an optimal prenatal environment for the developing foetus, improved pregnancy outcomes, and effective parental and community-wide interventions to

enhance child development. Another example is the engagement of parents in child-awareness programmes to facilitate their cognitive and socioemotional development. We have cited examples of specific family-based<sup>72</sup> and community-based<sup>73</sup> interventions that have been successfully used in LMICs. Such interventions involve the integration of anthropology, public health education and promotion, social and media science, and developmental paediatric research.

## Employment of new technologies

New technologies are enhancing research approaches in LMICs, presenting enormous potential to transform health-care delivery. The development of new mobile technologies for surveillance, assessment and treatment are particularly needed in LMICs, where mobile phone ownership is rapidly rising. Computerized interventions are already being used for the treatment of children with cerebral malaria and HIV. New and improved surgical techniques will be crucial for saving lives and altering atypical developmental trajectories. The miniaturization of diagnostic technologies that provide data for integrative risk maps, which can be integrated at the population level with genome distributions, will allow for effective population surveillance and public health intervention at a community-wide level.

## Implementation research

Research in LMICs cannot be divorced from the health systems and the cultural context in which the populations are situated. Research on the implementation of evidence-based prevention and intervention is needed. Scientifically sound interventions scaled up to the community and national level will require working with governmental and non-governmental partners to ensure sustainability. As already noted, significant advances have been made by 'task shifting' in resource-constrained settings in order to, for example, delegate health-care tasks to health workers with lower qualifications. Such strategies have shown especially promising results in dealing with mental health gaps, but can be effectively applied for the rehabilitative care of neurodisability in children<sup>74</sup>. Task-shifting strategies, however, need to be evaluated for approaches within dissemination and implementation science. It is only in this manner that we will achieve the UN moral and legal mandate of full childhood neurodevelopment as a basic human right for all.

1. UN General Assembly. *Political Declaration of the High-level Meeting of the General Assembly on the Prevention and Control of Non-communicable Diseases* [http://www.who.int/nmh/events/un\\_ncd\\_summit2011/political\\_declaration\\_en.pdf](http://www.who.int/nmh/events/un_ncd_summit2011/political_declaration_en.pdf) (United Nations, 2012).
2. Gluckman, P. *The Developmental Origins of Health and Disease* (Springer, 2006).
3. UNICEF. *The State of the World's Children: Children with Disabilities* (UNICEF, 2013).
4. United Nations. *Road Map Towards the Implementation of the United Nations Millennium Declaration*. Report No. A/56/326 (UN, 2002).
5. World Health Organization. *The World Malaria Report* (WHO, 2014).
6. Boivin, M. J. et al. Cognitive impairment after cerebral malaria in children: a prospective study. *Pediatrics* **119**, e360–e366 (2007).
7. John, C. C. et al. Cerebral malaria in children is associated with long-term cognitive impairment. *Pediatrics* **122**, e92–e99 (2008).
8. Boivin, M. J. et al. A year-long caregiver training program improves cognition in preschool Ugandan children with human immunodeficiency virus. *J. Pediatr.* **163**, 1409–1416 (2013).
9. Boivin, M. J. et al. A year-long caregiver training program to improve neurocognition in preschool Ugandan HIV-exposed children. *J. Dev. Behav. Pediatr.* **34**, 269–278 (2013).
10. Bangirana, P., Boivin, M. J. & Giordani, B. in *Neuropsychology of Children in Africa: Perspectives on Risk and Resilience* Vol. 1 (eds Boivin, M. J. & Giordani, B.) 277–298 (Springer, 2013).
11. Boivin, M. J. & Giordani, B. in *Cultural Neuroscience: Cultural Influences on Brain Function*. Vol. 178 (ed Chiao, J. Y.) 113–135 (Elsevier, 2009).
12. Brett, Z. H. et al. A neurogenetics approach to defining differential susceptibility to institutional care. *Int. J. Behav. Dev.* **39**, 150–160 (2015).
13. Croen, L. A. et al. Brain-derived neurotrophic factor and autism: maternal and infant peripheral blood levels in the Early Markers for Autism (EMA) study. *Autism Res.* **1**, 130–137 (2008).
14. UNAIDS. *UNAIDS Report on the Global AIDS Epidemic 2010* (UN, 2013).
15. Laughton, B., Cornell, M., Boivin, M. & Van Rie, A. Neurodevelopment in perinatally HIV-infected children: a concern for adolescence. *J. Int. AIDS Soc.* **16**, 18603 (2013).
16. Kvalsig, J. D., Taylor, M., Kauchali, S. & Chhagan, M. in *Neuropsychology of Children in Africa: Perspectives on Risk and Resilience* (eds Boivin, M. J. & Giordani, B.) Ch. 3, 37–67 (Springer, 2013).



17. Baral, S., Logie, C. H., Grosso, A., Wirtz, A. L. & Beyrer, C. Modified social ecological model: a tool to guide the assessment of the risks and risk contexts of HIV epidemics. *BMC Public Health* **13**, 482 (2013).
18. Noble, K. G. et al. Family income, parental education and brain structure in children and adolescents. *Nature Neurosci.* **18**, 773–778 (2015).
19. Reichman, N. E. & Kenney, G. M. Prenatal care, birth outcomes and newborn hospitalization costs: patterns among Hispanics in New Jersey. *Family Plan. Perspect.* **30**, 182–187 (1998).
20. Dewey, K. G. & Begum, K. Long-term consequences of stunting in early life. *Matern. Child Nutr.* **7** (Suppl 3), 5–18 (2011).
21. Kubota, T., Miyake, K., Hariya, N. & Mochizuki, K. Understanding the epigenetics of neurodevelopmental disorders and DOHaD. *J. Dev. Origins Health Dis.* **6**, 96–104 (2015).
22. Gregory, K. E. Microbiome aspects of perinatal and neonatal health. *J. Perinat. Neonatal Nurs.* **25**, 158–162 (2011).
23. National Institutes of Health. NIH-sponsored Study Identifies Superior Drug Regimen for Preventing Mother-to-Child HIV Transmission <http://www.nih.gov/news/health/nov2014/niad-17.htm> (NIH, 2014).
24. Sirois, P. A. et al. Safety of perinatal exposure to antiretroviral medications: developmental outcomes in infants. *Pediatr. Infect. Dis. J.* **32**, 648–655 (2013).
25. Guo, D., Shi, X., Song, D. & Zhang, L. Persistence of VRC01-resistant HIV-1 during antiretroviral therapy. *Sci. China. Life Sci.* **57**, 88–96 (2014).
26. Braithwaite, E. C., Kundakovic, M., Ramchandani, P. G., Murphy, S. E. & Champagne, F. A. Maternal prenatal depressive symptoms predict infant NR3C1 1F and BDNF IV DNA methylation. *Epigenetics* **10**, 408–417 (2015).
27. Murgatroyd, C., Quinn, J. P., Sharp, H. M., Pickles, A. & Hill, J. Effects of prenatal and postnatal depression, and maternal stroking, at the glucocorticoid receptor gene. *Transl. Psychiatry* **5**, e560 (2015).
28. Nemoda, Z. et al. Maternal depression is associated with DNA methylation changes in cord blood T lymphocytes and adult hippocampi. *Transl. Psychiatry* **5**, e545 (2015).
29. Perroud, N. et al. The Tutsi genocide and transgenerational transmission of maternal stress: epigenetics and biology of the HPA axis. *World J. Biol. Psychiatry* **15**, 334–345 (2014).
30. Nelson, R. Combating global health worker shortages: task shifting and sharing may provide one solution. *Am. J. Nursing* **112**, 17–18 (2012).
31. Swartz, L., Kilian, S., Twesigye, J., Attah, D. & Chiliza, B. Language, culture, and task shifting — an emerging challenge for global mental health. *Glob. Health Action* **7**, 23433 (2014).
32. Foltz, J. L. et al. An epidemiologic investigation of potential risk factors for nodding syndrome in Kitgum District, Uganda. *PLoS ONE* **8**, e66419 (2013).
33. Sejvar, J. J. et al. Clinical, neurological, and electrophysiological features of nodding syndrome in Kitgum, Uganda: an observational case series. *Lancet Neurol.* **12**, 166–174 (2013).
34. Couper, J. Prevalence of childhood disability in rural KwaZulu-Natal. *S. Afr. Med. J.* **92**, 549–552 (2002).
35. Idro, R. et al. Nodding syndrome in Ugandan children—clinical features, brain imaging and complications: a case series. *BMJ Open* **3**, e002540 (2013).
36. Koura, K. G. et al. Usefulness of child development assessments for low-resource settings in francophone Africa. *J. Dev. Behav. Pediatr.* **34**, 486–493 (2013).
37. Lozoff, B. et al. Long-lasting neural and behavioral effects of iron deficiency in infancy. *Nutr. Rev.* **64**, S34–S43 (2006).
38. Abubakar, A., Holding, P., Newton, C. R., van Baar, A. & van de Vijver, F. J. The role of weight for age and disease stage in poor psychomotor outcome of HIV-infected children in Kilifi, Kenya. *Dev. Med. Child Neurol.* **51**, 968–973 (2009).
39. Abubakar, A., Holding, P., Van de Vijver, F. J., Newton, C. & Van Baar, A. Children at risk for developmental delay can be recognised by stunting, being underweight, ill health, little maternal schooling or high gravidity. *J. Child Psychol. Psychiatry* **51**, 652–659 (2009).
40. Abubakar, A. et al. Socioeconomic status, anthropometric status, and psychomotor development of Kenyan children from resource-limited settings: a path-analytic study. *Early Hum. Dev.* **84**, 613–621 (2008).
41. Abubakar, A. in *Neuropsychology of Children in Africa: Perspectives on Risk and Resilience* (eds Boivin, M. J. & Giordani, B.) Ch. 9, 181–202 (Springer, 2013).
42. Sinclair, D., Abba, K., Grobler, L. & Sudarsanam, T. D. Nutritional supplements for people being treated for active tuberculosis. *Cochrane Database Syst. Rev.* **9**, CD006086 (2011).
43. Thankachan, P. et al. *Helicobacter pylori* infection does not influence the efficacy of iron and vitamin B12 fortification in marginally nourished Indian children. *Eur. J. Clin. Nutr.* **64**, 1101–1107 (2010).
44. Guerrant, R. L. et al. Mechanisms and impact of enteric infections. *Adv. Exp. Med. Biol.* **473**, 103–112 (1999).
45. Guerrant, R. L., Oria, R. B., Moore, S. R., Oria, M. O. & Lima, A. A. Malnutrition as an enteric infectious disease with long-term effects on child development. *Nutrition Rev.* **66**, 487–505 (2008).
46. Guerrant, R. L., Lima, A. A. & Davidson, F. Micronutrients and infection: interactions and implications with enteric and other infections and future priorities. *J. Infect. Dis.* **182**, S134–S138 (2000).
47. Guerrant, D. I. et al. Association of early childhood diarrhea and cryptosporidiosis with impaired physical fitness and cognitive function four-seven years later in a poor urban community in northeast Brazil. *Am. J. Trop. Med. Hyg.* **61**, 707–713 (1999).
48. Oria, R. B., Patrick, P. D., Blackman, J. A., Lima, A. A. & Guerrant, R. L. Role of apolipoprotein E4 in protecting children against early childhood diarrhea outcomes and implications for later development. *Med. Hypotheses* **68**, 1099–1107 (2007).
49. Oria, R. B. et al. APOE4 protects the cognitive development in children with heavy diarrhea burdens in Northeast Brazil. *Pediatr. Res.* **57**, 310–316 (2005).
50. Mitter, S. S. et al. Apolipoprotein E4 influences growth and cognitive responses to micronutrient supplementation in shantytown children from northeast Brazil. *Clinics* **67**, 11–18 (2012).
51. Cliff, J. et al. Konzo associated with war in Mozambique. *Trop. Med. Int. Health* **2**, 1068–1074 (1997).
52. Tylleskar, T. et al. Konzo: a distinct disease entity with selective upper motor neuron damage. *J. Neurol. Neurosurg. Psychiatry* **56**, 638–643 (1993).
53. Tylleskar, T., Legue, F. D., Peterson, S., Kpizungui, E. & Stecker, P. Konzo in the Central African Republic. *Neurology* **44**, 959–961 (1994).
54. Banea, J. P. et al. Survey of the konzo prevalence of village people and their nutrition in Kwilu District, Bandundu Province, DRC. *Afr. J. Food Sci.* **9**, 43–50 (2015).
55. Boivin, M. J. et al. Neuropsychological effects of konzo: a neuromotor disease associated with poorly processed cassava. *Pediatrics* **131**, e1231–e1239 (2013).
56. Bradbury, J. H., Cliff, J. & Denton, I. C. Uptake of wetting method in Africa to reduce cyanide poisoning and konzo from cassava. *Food Chem. Toxicol.* **49**, 539–542 (2010).
57. Warf, B. C. & the East African Neurosurgery Research Consortium. Pediatric hydrocephalus in East Africa: prevalence, causes, treatments, and strategies for the future. *World Neurosurg.* **73**, 296–300 (2010).
58. Kiwanuka, J. et al. The microbial spectrum of neonatal sepsis in Uganda: recovery of culturable bacteria in mother-infant pairs. *PLoS ONE* **8**, e72775 (2013).
59. Li, L. et al. Association of bacteria with hydrocephalus in Ugandan infants. *J. Neurosurg. Pediatr.* **7**, 73–87 (2011).
60. Schiff, S. J., Ranjeva, S. L., Sauer, T. D. & Warf, B. C. Rainfall drives hydrocephalus in East Africa. *J. Neurosurg. Pediatr.* **10**, 161–167 (2012).
61. Warf, B. C. Hydrocephalus in Uganda: the predominance of infectious origin and primary management with endoscopic third ventriculostomy. *J. Neurosurg.* **102**, 1–15 (2005).
62. Thaver, D. & Zaidi, A. K. Burden of neonatal infections in developing countries: a review of evidence from community-based studies. *Pediatric Infect. Dis. J.* **28**, S3–S9 (2009).
63. Warf, B. C., Dagi, A. R., Kaaya, B. N. & Schiff, S. J. Five-year survival and outcome of treatment for postinfectious hydrocephalus in Ugandan infants. *J. Neurosurg. Pediatr.* **8**, 502–508 (2011).
64. Warf, B. C. et al. Costs and benefits of neurosurgical intervention for infant hydrocephalus in sub-Saharan Africa. *J. Neurosurg. Pediatr.* **8**, 509–521 (2011).
65. Stone, S. S. & Warf, B. C. Combined endoscopic third ventriculostomy and choroid plexus cauterization as primary treatment for infant hydrocephalus: a prospective North American series. *J. Neurosurg. Pediatr.* **14**, 439–446 (2014).
66. Kulkarni, A. V. et al. Outcomes of CSF shunting in children: comparison of Hydrocephalus Clinical Research Network cohort with historical controls: clinical article. *J. Neurosurgery Pediatr.* **12**, 334–338 (2013).
67. Alwan, A. & Modell, B. Recommendations for introducing genetics services in developing countries. *Nature Rev. Genetics* **4**, 61–68 (2003).
68. DeBoer, M. D. et al. Early childhood growth failure and the developmental origins of adult disease: do enteric infections and malnutrition increase risk for the metabolic syndrome? *Nutrition Rev.* **70**, 642–653 (2012).
69. Tobin, E. W. et al. DNA methylation differences after exposure to prenatal famine are common and timing- and sex-specific. *Human Mol. Gen.* **18**, 4046–4053 (2009).
70. McCall, C. E., El Gazzar, M., Liu, T., Vachharajani, V. & Yoza, B. Epigenetics, bioenergetics, and microRNA coordinate gene-specific reprogramming during acute systemic inflammation. *J. Leukocyte Biol.* **90**, 439–446 (2011).
71. Fernald, L. C., Grantham-McGregor, S. M., Manandhar, D. S. & Costello, A. Salivary cortisol and heart rate in stunted and nonstunted Nepalese school child. *Eur. J. Clin. Nutrition* **57**, 1458–1465 (2003).
72. Elder, J. P. et al. Caregiver behavior change for child survival and development in low- and middle-income countries: an examination of the evidence. *J. Health Commun.* **19**, S25–S66 (2014).
73. Farnsworth, S. K. et al. Community engagement to enhance child survival and early development in low- and middle-income countries: an evidence review. *J. Health Commun.* **19**, S67–S88 (2014).
74. Mendenhall, E. et al. Acceptability and feasibility of using non-specialist health workers to deliver mental health care: stakeholder perceptions from the PRIME district sites in Ethiopia, India, Nepal, South Africa, and Uganda. *Soc. Sci. Med.* **118**, 33–42 (2014).


## ACKNOWLEDGEMENTS

The authors thank N. Anand, K. Michels, D. Silberberg and D. Krotoski for their editorial oversight, substantive input, guidance and support in this Review. I. Familiar-Lopez and H. Ruiseñor-Escudero provided help with Figure 1 and editorial comments for the final drafts and references as postdoctoral research associates for M.J.B. B. Giordani and V. Kutlesic provided advice and support to M. J. B. during the manuscript development process.

## COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests. Financial support for publication has been provided by the Fogarty International Center.

## ADDITIONAL INFORMATION

 This work is licensed under the Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0>

## REVIEW OPEN

# A focus on adolescence to reduce neurological, mental health and substance-use disability

Leslie L. Davidson<sup>1</sup>, Elena L. Grigorenko<sup>2</sup>, Michael J. Boivin<sup>3</sup>, Elizabeth Rapa<sup>4</sup> & Alan Stein<sup>4,5</sup>

Globally, there is a crucial need to prioritize research directed at reducing neurological, mental health and substance-use disorders in adolescence, which is a pivotal age for the development of self-control and regulation. In adolescence, behaviour optimally advances towards adaptive long-term goals and suppresses conflicting maladaptive short-lived urges to balance impulsivity, exploration and defiance, while establishing effective societal participation. When self-control fails to develop, violence, injury and neurological, mental health and substance-use disorders can result, further challenging the development of self-regulation and impeding the transition to a productive adulthood. Adolescent outcomes, positive and negative, arise from both a life-course perspective and within a socioecological framework. Little is known about the emergence of self-control and regulation in adolescents in low- and middle-income countries where enormous environmental threats are more common (for example, poverty, war, local conflicts, sex trafficking and slavery, early marriage and/or pregnancy, and the absence of adequate access to education) than in high-income countries and can threaten optimal neurodevelopment. Research must develop or adapt appropriate assessments of adolescent ability and disability, social inclusion and exclusion, normative development, and neurological, mental health and substance-use disorders. Socioecological challenges in low- and middle-income countries require innovative strategies to prevent mental health, neurological and substance-use disorders and develop effective interventions for adolescents at risk, especially those already living with these disorders and the consequent disability.

*Nature* 527, S161–S166 (19 November 2015), DOI: 10.1038/nature16030

This article has not been written or reviewed by *Nature* editors. *Nature* accepts no responsibility for the accuracy of the information provided.

Adolescent neurological, mental health and substance-use (NMS) disorders in low- and middle-income countries (LMICs) must be addressed to ensure optimal development of ‘human capital’ for the future<sup>1</sup>. Adolescents, defined by the World Health Organization (WHO) as those between 10 and 19 years of age, represent an estimated 1.2 billion of the world’s population<sup>2</sup>. During these years, cognitive development continues to unfold, socioemotional development advances dramatically, and a constellation of NMS disorders reaches a peak during adolescence and early adulthood<sup>3,4</sup> (Fig. 1). Research and practice in LMICs have focused on children under five years old, whereas adolescents are rarely prioritized.

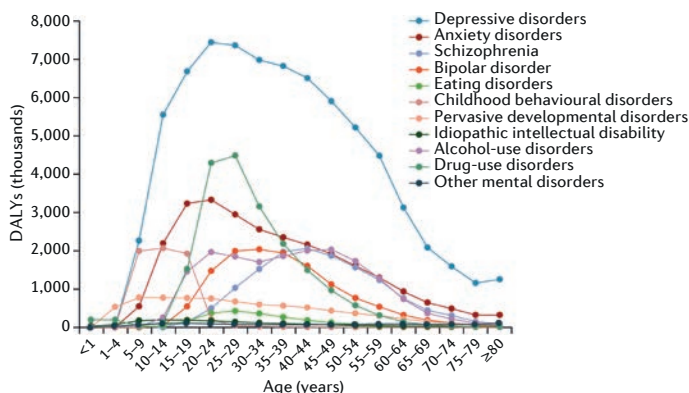
Indeed, more than 15% of disability-adjusted life years (DALYs) are attributable to adolescents and young adults (20–24 years old), with DALY rates in Africa 2.5 times that of high-income countries<sup>5</sup>. Globally, the primary causes of years lost due to disability (YLD) for adolescents include neuropsychiatric disorders (45%), unintentional injuries (12%) and infectious diseases (10%)<sup>5</sup>. Up to 20% of young adults have a disabling mental illness, and up to 50% of adult mental health disorders experience their onset in adolescence<sup>6</sup>. Adolescence has received more attention since being made a high priority by the United Nations programme UNICEF and other agencies<sup>1,7</sup>, facilitating opportunities for prevention of NMS disorders and support for adolescents already living

with these conditions (Box 1).

Adolescence is a crucial period of brain development that leads to increased self-regulation. However, concomitant impulsiveness can lead to risky behaviours that result in impaired cognitive or emotional development, lifelong disability and even death. These behaviours and challenging environmental exposures often interact over time, compounding their effects. For example, early initiation of substance misuse might lead to impaired cognitive and affective development, later addiction, brain injury and other health-related disorders such as HIV/AIDS<sup>8</sup>.

The impact of NMS disorders on adolescents is best understood from a life-course perspective. In this Review, we highlight only the causes of NMS disorders that tend to first appear during adolescence. We consider risk and resilience through a socioecological model in which an adolescent with a particular genome undergoes physical and hormonal transformations within the context of a family, peers, school, work, community and culture all of which can be instrumental in determining adolescent NMS outcomes. Few studies have been conducted using parallel methods across high-income countries and LMICs, and since it is not clear to what extent findings in high-income countries can be generalized to LMICs, we suggest that there is a need for comparative research using parallel methods. This Review focuses on research priorities to reduce the most important threats to

<sup>1</sup>Departments of Epidemiology and Pediatrics, Mailman School of Public Health and College of Physicians and Surgeons, Columbia University, 722 West 168 Street Room 1613, New York, New York 10032, USA. <sup>2</sup>Child Study Center, Department of Epidemiology and Public Health, Department of Psychology, Yale University, 230 South Frontage Road, New Haven, Connecticut 06519, USA. <sup>3</sup>Department of Psychiatry and Department of Neurology and Ophthalmology, College of Osteopathic Medicine, 965 Fee Road, Room A227, Michigan State University, East Lansing, Michigan 48824, USA. <sup>4</sup>Department of Psychiatry, University of Oxford, Oxford, OX3 7JX, UK. <sup>5</sup>MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa. Correspondence should be addressed to L. L. D. e-mail: LLD1@columbia.edu.



**Figure 1** | Disability-adjusted life years (DALYs) for each neurological, mental health and substance-use (NMS) disorder in 2010 by age\*. Note the rise of NMS disorders in late childhood and adolescence, particularly depression, anxiety, alcohol and other drug-use disorders. Reprinted with permission from ref. 4.

adolescents in LMICs and contains suggestions to improve methodology and a brief list of high-impact research priorities (Box 2). We also suggest a range of broader research opportunities, priorities and challenges, including research capacity building (Supplementary Table 1).

## NEURODEVELOPMENT AND NMS DISORDERS

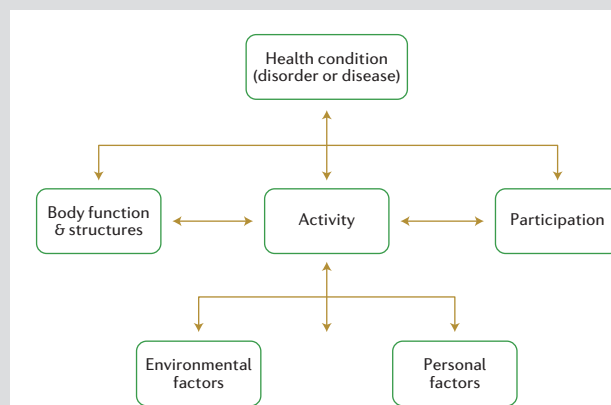
Adolescence is a time of profound change as young adults cope with increasing independence and the growing importance of social and sexual relationships, while simultaneously developing and exercising self-control. It is unknown whether hormonal, cognitive, affective and societal transitions from childhood to adolescence occur in a similar time frame and sequence in LMICs as they do in high-income countries, or whether any differences in timing or sequencing affect neural development. The likelihood that an adolescent will develop a specific NMS-related disability depends to some extent on the balance between higher order cognitive function (known as executive control) and environmental pressures, societal expectations and impulsivity, and the way in which the individual processes emotional, social and behavioural cues<sup>9-12</sup> (Fig. 2). The impact of adolescent NMS disorders can be exacerbated by early exposure to risk factors that are more prevalent in LMICs (for example, malnutrition, infectious disease, abuse, neglect, internal conflict or war, early pregnancy and marriage, and adolescent labour). Consequently, adolescents in LMICs may be at a higher risk than their counterparts in high-income countries.

The recent Well-Being of Adolescents in Vulnerable Environments study used identical methods to study adolescents from disadvantaged areas in five cities: Baltimore, Shanghai, Delhi, Ibadan and Johannesburg. Neighbourhood physical and social exposures were associated with adolescent disorders in LMICs and high-income countries, although patterns and degree of association varied. The authors concluded that “for young people growing up in poverty, residency in a high-income country may matter far less than the immediate social contexts within which they develop and grow”<sup>13</sup>. There is also evidence<sup>14</sup> that in settings where socioeconomic disparities are great that depression is especially high compared with places where poverty is almost universal. Such findings illustrate the need for longitudinal and cross-sectional multi-country studies using comparable methods across differing levels of economic and social development.

## PREVENTION OF NMS DISABILITY

An overview of research priorities for prevention of NMS disability is presented in Supplementary Table 1. Adolescents and young adults experience the highest amounts of violence, both towards self (self-harm and suicide) and others (partner violence, delinquent behaviour, criminal acts and war-related trauma)<sup>2</sup>. In addition, adolescents are at increased risk of a range of health problems that can lead to NMS disability

## BOX 1 | HUMAN RIGHTS AND ADOLESCENT RESEARCH



Two major international conventions and a new classification system transformed opportunities to support optimal development for children and adolescents living with neurological, mental health and substance-use disorders and for the primary prevention of disabling disorders. The Convention on the Rights of the Child (which includes adolescents) and the Convention on the Rights of People with Disabilities hold nations responsible for ensuring optimal developmental outcomes for children and adolescents and guarantee rights for those living with disabilities. These rights also commit countries to support adolescents in social inclusion and participation in society, and reduce stigma.

The International Classification of Functioning, Disability and Health for Children and Youth (ICF-CY) takes account of the interaction between the biology of a given individual and the specific external challenges or supports affecting an individual's ability to function. It also acknowledges societal structures (environmental factors) that decrease or increase barriers to participation and inclusion<sup>47</sup>. The ICF-CY, outlined in the Figure, relies on concepts of disability rather than disorder-specific pathology or morbidity and incorporates a socioecological framework. This highlights the importance of environmental context alongside genetic risk as a child develops through adolescence to adulthood.

such as HIV/AIDS, meningitis, and trauma due to civil and interpersonal violence, as well as unintentional injury such as road traffic accidents<sup>5</sup>.

## Gender inequality, education, pregnancy and violence

Increases in gender inequality during adolescence are evident in many LMICs, with a higher proportion of boys than girls completing secondary school. This directly results in a substantial gender gap in education and literacy rates<sup>2</sup> and indirectly in differential opportunities for the development of executive function, self-regulation and employability. Intimate partner violence (IPV) against women also peaks during adolescence and is a known risk factor for acquiring HIV as well as brain trauma, post traumatic stress disorder (PTSD) and depression. The WHO multicountry study found that in women aged 15–24 who have ever had a romantic or sexual partner, the prevalence of IPV experience before the age of 15 ranged from 19% to 64%, with most sites reporting that this occurred in more than 50% of participants. There is little information on NMS outcomes in adolescent women experiencing IPV in LMICs, although in high-income countries, PTSD, depression and substance misuse are frequently associated with IPV<sup>15</sup>. Girls are at higher risk of HIV and AIDS, whereas boys are more at risk of civil and interpersonal violence<sup>16</sup>.



Adolescent mothers in high-income countries are particularly at risk of developing depression: prenatal, postnatal and, in one study, long-term depression<sup>17</sup>, but we found no research on the impact of adolescent pregnancy on depression in LMICs. The impact of adolescent pregnancy on maternal cognitive development has not been studied in any context and represents a global research priority. This is especially relevant in LMICs where 20% of mothers have had their first child by the age of 18 and more than 20% of girls are married before age 18, despite a ban based on international conventions<sup>16</sup>.

### Community and conflict-related violence and injury

Adolescence is a dynamic time of biological (brain and sexual maturation) and social development (differentiating self from others and appraising the self by forming self-control, self-esteem and self-efficacy). Adolescents are concomitantly more likely to break social rules and experience aggression towards self and others<sup>9–11</sup>. The peak age of both committing and receiving violent acts is between 14 and 19 years old — adolescent males are more frequently the perpetrators and victims than are females<sup>2</sup>.

In wars, despite international agreements and conventions to the contrary, child soldiers are pressed into service at a young age, experiencing disrupted social and emotional development and interrupted schooling. These adolescents are at risk of physical trauma with lasting disability, PTSD and addiction<sup>18</sup>. Similar outcomes are found from the impact of post-conflict displacement of adolescents, who are often separated from their family and community<sup>19</sup>. However, there is evidence that negative impacts can be mitigated by stable and supportive social environments. A study of 880 Bosnian adolescents demonstrated the need for better methodological approaches to unpack the multitiered pathways that lead to trauma in order to develop better interventions<sup>20</sup>. A review of interventions providing mental health support to conflict-affected children and adolescents found that most were school-based programmes<sup>21</sup>. There were few family- or community-based programmes and only two multilevel programmes. A systematic review of interventions for refugee children found eight studies in LMICs (seven in refugee camps) that showed that treatments, including cognitive behavioural therapy (CBT) and narrative exposure therapy, were successful in reducing psychological problems such as depression and PTSD<sup>22</sup>.

There is limited evidence that interventions developed in high-income countries can be transplanted successfully<sup>22,23</sup>. It has been argued<sup>24</sup> that the most effective strategies for preventing violence should promote mental health and be delivered to all young people, thus avoiding stigmatization and attracting broad community support. Yet, the evidence base for effective programmes to prevent violence in LMICs is almost non-existent<sup>3,6</sup>. A recent review of interventions to support street children found 12 studies in high-income countries, but did not find any adequately robust interventions in LMICs despite the plethora of programmes that are being implemented<sup>25,26</sup>.

Road traffic accidents and other transport-related injuries are the leading cause of brain trauma and spinal cord injury, death and disability, disproportionately affecting young adults and accounting for about 5% of all DALYs, which is only exceeded by unipolar depression<sup>5</sup>.

### Mental health and substance use

Difficulties in the development of executive control in adolescence can lead to a lack of balance in the regulation of cognition, emotions and behaviour when dealing with intrusive negative thoughts and feelings. This may result in depression or anxiety. Such affective disorders often have their onset in adolescence (Fig. 1) when these disorders are more likely to become severe and disabling<sup>9</sup>. The prevalence of depression in adolescent girls is substantially higher globally than that in boys<sup>5</sup>. There is some evidence from high-income countries that early intervention to reduce the duration of first episodes of depression can reduce later recurrence<sup>27</sup>; this needs to be evaluated in LMICs.

The first onset of schizophrenia often arises in adolescence or

early adult life (Fig. 1). Early identification of psychosis by a range of non-specialists followed by intervention may mitigate the severity of the disease and the need for hospitalization<sup>28</sup>. Family approaches have also shown promise for young adults at high risk<sup>29</sup>. However, differences between high-income countries and LMICs in risk factors for, and the effective prevention and treatment of, psychoses are poorly understood<sup>30</sup>.

Adolescents can be at increased risk of developing substance-use disorders owing to poor self-control and impulsiveness. Although not well studied in LMICs, substance-misuse is often initiated in adolescence and accounts for a substantial proportion of the disability burden faced by adolescents<sup>31</sup>. It also contributes to other major causes of disability such as unintentional injuries and violence<sup>5</sup>. Alcohol use, particularly binge drinking, has been shown to inhibit neuronal development in adolescence<sup>32</sup> and has a greater impact on motor and executive impairment in adolescents than in adults, thus conferring greater risk of injury or risky behaviour in adolescents<sup>33</sup>.

In Russia, excessive alcohol consumption is a major cause of premature death<sup>34</sup> and is associated with early initiation and frequent alcohol consumption in adolescence<sup>35</sup>. Evidence-based interventions to delay onset of drinking and reduce binge drinking in adolescents in LMICs are a high priority. A related priority calls for the inclusion of measures of depression, violence and sexually transmitted diseases in research aimed at diminishing the impact of either early initiation of drinking or of binge drinking in adolescence.

### Infectious diseases

Vaccine-preventable neurological disorders may result in cognitive disability, epilepsy, motor disorders, and hearing and/or vision loss. For example, meningitis A is prevalent in sub-Saharan Africa and targets adolescents and young adults, causing death, cognitive disability and hearing loss. WHO and PATH have developed a new vaccine against meningitis A that has been introduced in a number of African countries. More than 100 million doses have been delivered to people between 1 and 29 years of age, contributing to the lowest incidence of the disease in 10 years<sup>36</sup>. The use of implementation science to further scale-up this intervention would eliminate a major cause of hearing loss and cognitive disability in African adolescents.

Untreated HIV infection is associated with disabling cognitive impairment, depression and behavioural disorders in adolescence<sup>37</sup>. Following the scale-up of antiretroviral (ARV) therapy, survival of perinatally infected children in sub-Saharan Africa has dramatically improved. The population of adolescents with HIV is now estimated at 2 million with more than 90% of those living in sub-Saharan Africa. Young adults infected with HIV are at increased risk of developmental and neuropsychological disturbances, which seriously undermine academic and social achievement<sup>37,38</sup>. A study in South Africa evaluated a community participatory approach to adapt a family-based intervention (originally developed in the United States) to promote mental health awareness in adolescents receiving HIV ARV treatment<sup>37</sup>. Short-term results showed improvements in mental health, behaviour and adherence to the drug regimen with a decrease in stigma experience<sup>37</sup>. Peer-led programmes are now part of many large-scale initiatives to reduce HIV risk in young adults, and a systematic review found that they are effective in increasing knowledge, particularly of HIV prevention approaches and transmission routes and in increasing condom use in LMICs, but evidence of changes in sexual behaviour and STI rates are not conclusive<sup>39</sup>. Research to determine the effectiveness, generalizability and the long-term impact of such interventions to support adolescent development is a priority.

### Research in developing educational initiatives

An important challenge to adolescents' development in LMICs is the low rate of secondary education completion<sup>2</sup>. A review of competencies needed by vulnerable young adults as a result of war, homelessness and child labour concluded that these are not the same as those

conventionally required in the education of adolescents in high-income countries, but posits that the process by which they are acquired is universal<sup>18</sup>. Research is needed for the development, evaluation and delivery of necessary competencies in secondary education in LMICs. Such innovative approaches would add to standard interventions to improve health, mental health and social outcomes. Such programmes would focus not only on the acquisition of skills and knowledge, but also on the development of cognitive and motivational skills that are central to the emergence of self-regulated learning.

## ADOLESCENTS WITH DISABILITIES RESEARCH

There are few studies that evaluate the approaches that support LMIC adolescents living with disabilities. A systematic review of 22 school-based interventions in LMICs found promising support for mental health promotion and some evidence that interventions to prevent and to treat mental health disorders are effective<sup>40</sup>. The study identified a series of research gaps that need to be filled to understand how to further evaluate and bring to scale these school-based interventions. These include the need to focus on current stressors such as social exclusion and domestic violence, and to test interventions as part of a multilayered societal programme with an emphasis on task shifting (use of trained mid-level researchers in place of clinical specialists).

A priority area for research is to develop and evaluate programmes to ensure that adolescents living with an NMS disability are not excluded from access to social, educational and health services that would reduce secondary and tertiary development of disability and ensure optimal development. Interventions that increase independence must also be developed and tested to protect adolescents from violence, and to ensure educational parity and an effective transition to adult services<sup>41</sup>.

## METHODOLOGICAL ISSUES

It is often challenging to include adolescents in research. With the increasing independence of adolescents, their parents may not have accurate information regarding exposures and even outcomes. The home may no longer be a feasible setting for any study that requires either a physical examination or biological samples. However, recent studies have shown that providing support to carers and communities that assist vulnerable adolescents facing HIV infection can mitigate obstacles in reaching such groups<sup>37,42</sup>. Studies based in schools are often used, although such approaches may only work for younger (10–13 year olds) populations because most LMICs adolescents are no longer in school. According to UNICEF<sup>2</sup> only 61% of boys and 49% of girls entered secondary school in LMICs. One successful approach to reach vulnerable or marginalized urban adolescents is respondent-driven sampling<sup>43</sup>.

## Disaggregation of age group

Researchers should collect data to allow disaggregation by age to facilitate, at low additional cost, age-specific estimates of exposures and outcomes. Often research data on adolescents are combined with either children or adults<sup>3</sup>. A systematic review, including studies from LMICs and high-income countries, demonstrated that the highest prevalence of intellectual disability was among children and adolescents compared with adults, and within LMICs compared with high-income countries<sup>44</sup>. However, it was impossible to disaggregate adolescents from children or adults. As an example of the importance of disaggregation<sup>45</sup>, a peer support-group intervention to reduce the impact of stigma in those living with epilepsy in Zambia found a significant effect in adolescents, but not in adults.

## Promoting and funding longitudinal studies

Major birth cohort studies in high-income countries have provided information on risks and resilience across the life course, but these are not comparable for LMICs. There are several long-standing birth cohort studies in LMICs such as those in Guatemala, India, Brazil, the Philippines and South Africa<sup>46</sup>. The cohorts in such studies should be supported over the life course and the studies should be further developed to

## BOX 2 | RECOMMENDATIONS FOR RESEARCH

Recommendations for methodological priorities for research to reduce neurological, mental health and substance-use (NMS) disorders in adolescents in low- and middle-income countries (LMICs).

- Disaggregate adolescent age group in both child and adult studies
- Promote, develop and fund longitudinal cohort studies in LMICs, including observational and follow-up of interventions over the long term
- Initiate cross-nation studies of interventions to promote resilience in adolescents
- Include constellations of exposures and outcomes in studies that involve adolescents and avoid focusing on a single exposure or outcome
- Use methodological approaches to explore multisectoral and multilevel pathways leading to NMS disorders in traumatized young adults
- Include measures of disability and participation in studies of adolescents
- Use new screening measures that include adolescence developed for national studies (UNICEF and Washington Group on Disability Statistics)
- Link development of research priorities to emerging findings from national studies
- Train mid-level researchers (nurses, teachers and others) to conduct research, including assessment of NMS disorders where shown to be effective

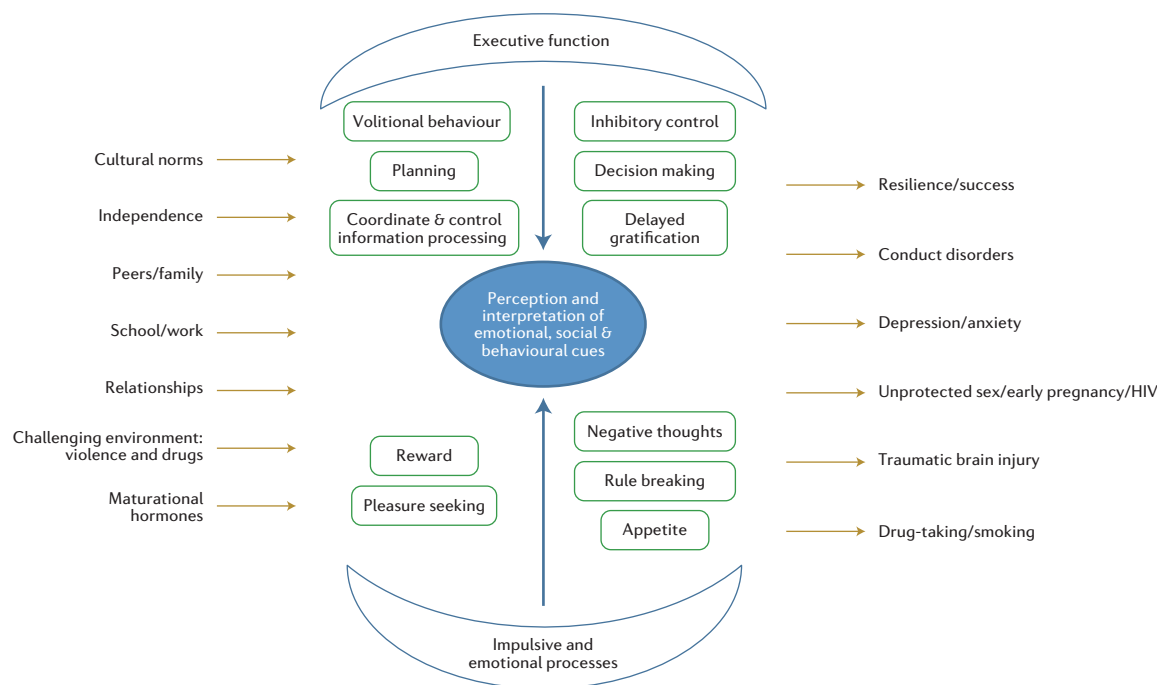
High impact priorities for research to reduce NMS disorders in adolescents in LMICs.

- Develop and evaluate strategies to provide mental health interventions for adolescents in community and school settings and address gaps
- Develop an evidence base for interventions designed to prevent violence in young adults
- Investigate the impact of early child bearing on adolescent mothers on development of executive function
- Develop innovative approaches to provide appropriate secondary education for all adolescents
- Develop innovative approaches to ensure inclusive education for those living with disabilities
- Develop and evaluate approaches to protect those living with disabilities from violence and to ensure participation in their communities and cultures
- Support evaluation of interventions to support the transition between adolescent and adult life for those living with NMS disabilities in LMICs

address the research questions that are central to preventing or reducing the impact of adolescent NMS disorders and their consequences.

## Broadening outcomes measures

Since there is a constellation of exposures that put adolescents at risk of NMS disorders (many of which may simultaneously be NMS outcomes of other exposures), research into NMS disorders in adolescence should, ideally, not focus on a single exposure or outcome, but should measure multiple factors, whether exposure or outcome. In order to avoid too narrow a sectoral focus, researchers should include a summary measure of disability in addition to measures specific to their study question and consider environmental exposures across education, social development and health, using the framework of the International Classification of Functioning, Disability and Health for Children and



**Figure 2** | The interaction between ‘top down’ executive function and ‘bottom up’ impulsive and emotional processes in regulating social, cultural and biological challenges of adolescent development with a range of outcomes<sup>9–12</sup>.

Youth (ICF-CY)<sup>47</sup>. To accomplish this cross-cultural age-appropriate measures of disability, environmental risk and social participation need to be developed and validated to assess the balance between the developing abilities of the adolescent within the context of the barriers faced in their own communities and culture.

Developing research capacity to measure adolescent NMS outcomes beyond mortality and conventional morbidity measures is essential. UNICEF and the Washington Group on Disability Statistics collaborated to extend internationally appropriate screening measures of disability to include the adolescents. The WHO has developed a child and adolescent form of the brief WHO Disability Assessment Schedule<sup>48</sup>, which is designed to provide information regarding the function of young adults living with disabilities, including NMS disorders. This schedule uses the ICF-CY to provide information on external supports and barriers, and societal participation. It is a priority to include these measures in national surveys and in research studies, to provide both comparative data and population estimates of adolescent disability, which can inform national strategies.

## CONCLUSION

Because of the range of environmental challenges that are more common in LMICs, adolescents are at particular risk of developing neurocognitive deficits and disabilities, and mental health problems that limit their reasoning abilities, life-management skills and employability. These are often related to the poor development of executive function or socioemotional development caused by traumatic brain injury due to war or interpersonal conflict, domestic violence or abuse, substance use, or infectious diseases. These largely unmeasured NMS disorders cannot be prevented or treated by an under-resourced health and educational infrastructure. Therefore, it is imperative to develop neurodevelopmental and psychiatric screening strategies to identify and assess these young adults. These will make it possible to evaluate the effectiveness of prevention and early intervention programmes for NMS disorders. Along with screening and clinical assessment for neurocognitive and neuropsychiatric disorders in adolescents at high risk, it is crucial to provide comprehensive access to effective rehabilitative neurocognitive and psychosocial interventions. A scaffold of universal

supports for those at high risk of NMS and with disability will also require evaluation of more intensive interventions for those with significant disorders.

The recent extension of screening measures to include NMS disability in adolescents for use in national or regional surveys provides new opportunities for prioritizing research and directing funding streams in LMICs, and provide the data on which programmes and policies can be based.

Adolescence is not simply a phase between childhood and adulthood. Research to prevent and ameliorate NMS disorders in LMICs must bring an understanding of the interlocking neurobiological and social factors that challenge young adults, and recognize that the patterns of risk and resilience may differ from those in high-income countries. Research must investigate approaches to prevent toxic exposures such as community violence, sex trafficking, drug addiction, poverty and the harm caused by war. Research should develop and evaluate interventions to prevent the emergence of disability from NMS disorders in adolescents in LMICs, as well as to develop and test interventions to support adolescents living with disability to ensure appropriate schooling, employment and the opportunity to become productive members of society, with a rewarding quality of life.

1. World Bank. *World Bank Development Report 2007* (World Bank, 2007).
2. UNICEF. *Progress for Children: A Report Card on Adolescents* (UNICEF, 2012).
3. Patel, V., Flisher, A. J., Nikapota, A. & Malhotra, S. Promoting child and adolescent mental health in low and middle income countries. *J. Child Psychol. Psychiatry* **49**, 313–334 (2008).
4. Whiteford, H. A. et al. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *Lancet* **382**, 1575–1586 (2013).
5. Gore, F. M. et al. Global burden of disease in young people aged 10–24 years: a systematic analysis. *Lancet* **377**, 2093–2102 (2011).
6. Belfer, M. L. Child and adolescent mental disorders: the magnitude of the problem across the globe. *J. Child Psychol. Psychiatry* **49**, 226–236 (2008).
7. UNICEF. *The State of the World's Children: Children with Disabilities* (UNICEF, 2013).
8. Sawyer, S. M. et al. Adolescence: a foundation for future health. *Lancet* **379**, 1630–1640 (2012).
9. Paus, T., Keshavan, M. & Giedd, J. N. Why do many psychiatric disorders emerge during adolescence? *Nature Rev. Neurosci.* **9**, 947–957 (2008).
10. Casey, B. J., Jones, R. M. & Hare, T. A. The adolescent brain. *Ann. New York Acad. Sci.* **1124**, 111–126 (2008).



11. Steinberg, L. Cognitive and affective development in adolescence. *Trends Cog. Sci.* **9**, 69–74 (2005).
12. Blakemore, S. J. The social brain in adolescence. *Nature Rev. Neurosci.* **9**, 267–277 (2008).
13. Blum, R. W. Distressed communities as a breeding ground for noncommunicable conditions. *J. Adolesc. Health* **55**, S4–S5 (2014).
14. Lorant, V. et al. Socioeconomic inequalities in depression: a meta-analysis. *Am. J. Epidemiol.* **157**, 98–112 (2003).
15. Stöckl, H. M. L., Pallitto, C., Garcia-Moreno, C. & WHO Multi-Country Study Team. Intimate partner violence among adolescents and young women: prevalence and associated factors in nine countries: a cross-sectional study. *BMC Public Health* **14**, 751 (2014).
16. UNICEF. *Ending Child Marriage: Progress and Prospects* (UNICEF, 2012).
17. Hodgkinson, S., Beers, L., Southammosane, C. & Lewin, A. Addressing the mental health needs of pregnant and parenting adolescents. *Pediatrics* **133**, 114–122 (2014).
18. Grigorenko, E. L. & O'Keefe, P. In *Culture and Competence* 22–53 (American Psychological Assoc., 2004).
19. Reed, R. V., Fazel, M., Jones, L., Panter-Brick, C. & Stein, A. Mental health of displaced and refugee children resettled in low-income and middle-income countries: risk and protective factors. *Lancet* **379**, 250–265 (2012).
20. Layne, C. M. et al. Unpacking trauma exposure risk factors and differential pathways of influence: predicting postwar mental distress in Bosnian adolescents. *Child Dev.* **81**, 1053–1076 (2010).
21. Betancourt, T. S., Meyers-Oski, S. E., Charrow, A. P. & Tol, W. A. Interventions for children affected by war: an ecological perspective on psychosocial support and mental health care. *Harv. Rev. Psych.* **21**, 70–91 (2013).
22. Tyrer, R. A. & Fazel, M. School and community-based interventions for refugee and asylum seeking children: a systematic review. *PLoS ONE* **9**, e89359 (2014).
23. Wuermli, A. J., Tubbs, C. C., Petersen, A. C. & Aber, J. L. Children and youth in low- and middle-income countries: toward an integrated developmental and intervention science. *Child Develop. Perspect.* **9**, 61–66 (2015).
24. Offord, D. R. Selection of levels of prevention. *Addict. Behav.* **25**, 833–842 (2000).
25. Patel, V. et al. Treatment and prevention of mental disorders in low-income and middle-income countries. *Lancet* **370**, 991–1005 (2007).
26. Coren, E. et al. Interventions for promoting reintegration and reducing harmful behaviour and lifestyles in street-connected children and young people. *Evid. Based Child Health* **8**, 1140–1272 (2013).
27. Patton, G. C. et al. The prognosis of common mental disorders in adolescents: a 14-year prospective cohort study. *Lancet* **383**, 1404–1411 (2014).
28. McFarlane, W. R. et al. Reduction in incidence of hospitalizations for psychotic episodes through early identification and intervention. *Psych. Serv.* **65**, 1194–1200 (2014).
29. Miklowitz, D. J. et al. Family-focused treatment for adolescents and young adults at high risk for psychosis: results of a randomized trial. *J. Am. Acad. Child Adolesc. Psychiatry* **53**, 848–858 (2014).
30. Morgan, C. et al. Searching for psychosis: INTREPID (1): systems for detecting untreated and first-episode cases of psychosis in diverse settings. *Social Psychiatry Psychiatric Epidemiol.* **50**, 879–893 (2015).
31. Toumbourou, J. W. et al. Interventions to reduce harm associated with adolescent substance use. *Lancet* **369**, 1391–1401 (2007).
32. Ehlers, C. L. & Criado, J. R. Adolescent ethanol exposure: does it produce long-lasting electrophysiological effects? *Alcohol* **44**, 27–37 (2010).
33. Matthews, D. B. Adolescence and alcohol: recent advances in understanding the impact of alcohol use during a critical developmental window. *Alcohol* **44**, 1–2 (2010).
34. Zaridze, D. et al. Alcohol and mortality in Russia: prospective observational study of 151,000 adults. *Lancet* **383**, 1465–1473 (2014).
35. Verho, A., Laatikainen, T., Vartiainen, E. & Puska, P. Changes in alcohol behaviour among adolescents in North-West Russia between 1995 and 2004. *J. Environ. Public Health* **2012**, 736249 (2012).
36. World Health Organization. *Meningitis A Vaccine Now Recommended in Routine Immunization Schedules* <http://www.afro.who.int/en/media-centre/afro-feature/item/7312-meningitis-a-vaccine-now-recommended-in-routine-immunization-schedules.html> (WHO, 2015).
37. Bhana, A. et al. The VUKA family program: piloting a family-based psychosocial intervention to promote health and mental health among HIV infected early adolescents in South Africa. *AIDS Care* **26**, 1–11 (2014).
38. Hazra, R., Siberry, G. K. & Mofenson, L. M. Growing up with HIV: children, adolescents, and young adults with perinatally acquired HIV infection. *Ann. Review Med.* **61**, 169–185 (2010).
39. Maticka-Tyndale, E. & Barnett, J. P. Peer-led interventions to reduce HIV risk of youth: a review. *Eval. Prog. Planning* **33**, 98–112 (2010).
40. Fazel, M., P. V., Thomas S. & Tol W. Mental health interventions in schools in low-income and middle-income countries. *Lancet Psychiatry* **5**, 388–398 (2014).
41. Colver, A. F. et al. Study protocol: longitudinal study of the transition of young people with complex health needs from child to adult health services. *BMC Public Health* **13**, 675 (2013).
42. Casale, M. et al. Direct and indirect effects of caregiver social support on adolescent psychological outcomes in two South African AIDS-affected communities. *Am. J. Comm. Psychol.* **55**, 336–346 (2015).
43. Olumide, A. O. et al. Predictors of substance use among vulnerable adolescents in five cities: findings from the well-being of adolescents in vulnerable environments study. *J. Adolesc. Health* **55**, S39–S47 (2014).
44. Maulik, P. K., Mascarenhas, M. N., Mathers, C. D., Dua, T. & Saxena, S. Prevalence of intellectual disability: a meta-analysis of population-based studies. *Res. Develop. Disabilities* **32**, 419–436 (2011).
45. Elafros, M. A. et al. Peer support groups as an intervention to decrease epilepsy-associated stigma. *Epilepsy Behav.* **27**, 188–192 (2013).
46. Richter, L. M. et al. Cohort profile: the consortium of health-orientated research in transitioning societies. *Int. J. Epidemiol.* **41**, 621–626 (2012).
47. World Health Organization. *The International Classification of Functioning, Disability and Health. Children and Youth Version* (WHO, 2007).
48. World Health Organization. *WHO Disability Assessment Schedule 2.0 WHODAS 2.0* <http://www.who.int/classifications/icf/whodasii/en/index6.html> (WHO, 2015).

#### SUPPLEMENTARY INFORMATION

Is linked to the online version of this paper at: <http://dx.doi.org/10.1038/nature16030>.


#### ACKNOWLEDGEMENTS

We would like to recognize the assistance of A. Shankar, E. Netsi, V. Kutlesic, N. Anand, and the editors of this supplement in writing this Review.

#### COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests. Financial support for publication has been provided by the Fogarty International Center.

#### ADDITIONAL INFORMATION

 This work is licensed under the Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0>

REVIEW **OPEN**

# Nervous system disorders across the life course in resource-limited settings

Gretchen L. Birbeck<sup>1,2</sup>, Ana-Claire Meyer<sup>3,4</sup> & Adesola Ogunniyi<sup>5</sup>

The resiliency of the adult nervous system is markedly affected by the environment and the circumstances during infant and child development. As such, adults in resource-limited settings who may have experienced early deprivation are particularly vulnerable to subsequent neurological disorders. Adult populations in countries with relatively recent advances in economic development may still have a higher susceptibility to neurological illness or injury that is reflective of the socioeconomic environment that was present during that population's infancy and childhood. Brain and peripheral nervous system research conducted over the past decade in resource-limited settings has led to an impressive and growing body of knowledge that informs our understanding of neurological function and dysfunction, independent of geography. Neurological conditions feature prominently in the burgeoning epidemic of non-communicable diseases facing low- and middle-income countries. Neurological research in these countries is needed to address this burden of disease. Although the burden of more prevalent and severe neurological disease poses public health and clinical challenges in settings with limited neurological expertise, the same factors, along with genetic heterogeneity and the relative absence of ingrained clinical care practices, offer circumstances well-suited for the conduct of crucial future research that is globally relevant.

*Nature* 527, S167–S171 (19 November 2015), DOI: 10.1038/nature16031

This article has not been written or reviewed by *Nature* editors. *Nature* accepts no responsibility for the accuracy of the information provided.

Neurological disorders represent a significant proportion of the burden of disease among adults in low- and middle-income countries (LMICs). Stroke, epilepsy and dementia rank among the highest causes of death and disability, and often affect working-age adults. High rates of premature death and disability undoubtedly constrain societies that are striving to advance human development through improvements in health and a more stable economy. Research aimed at elucidating the pathophysiology and epidemiology of adult nervous system disorders across a broad spectrum of settings has advanced our capacity to develop interventions at the population and individual level. Insights gained from developed countries have provided remarkable evidence that deprivation in early life imparts additional health-related vulnerabilities that extend into adulthood<sup>1</sup>. This added susceptibility is an important and generally unmeasured factor that could account for the poor health status among LMICs adults. The burden of non-communicable diseases continues to escalate in LMICs that are still struggling with infectious diseases and maternal health challenges; the need to identify opportunities for optimizing health in older populations is increasingly urgent. In this Review, we detail an overview of some of the insights gained from research in diverse academic, laboratory and community settings and consider what knowledge can be gained from investments in future endeavours.

## EARLY EXPOSURE AND THE ADULT NERVOUS SYSTEM

Neurological disorders of infancy, childhood and adolescence are discussed elsewhere in this collection (see page S161), but if survived,

neurological injury in early life may result in adult disability, vulnerability to disease and premature mortality. Therefore, it is crucial to integrate a lifespan approach as we examine what we currently know, as well as research priorities for the future. Central nervous system injuries, infections and infestations during childhood that result in epilepsy, behavioural disorders and intellectual disabilities lay the foundation for a compromised adulthood. Furthermore, the life course and exposures of childhood determine adult neurological health, with childhood deprivation and adolescent injury among the primary risk factors for adult neurological disorders in LMICs. For example, childhood risk factors for epilepsy in sub-Saharan Africa include labour and delivery complications and lower maternal education levels<sup>2</sup>. Data from 1921 and 1936 birth cohorts in Aberdeen, UK, suggest that early nutritional deprivation predisposes people to dementia in later life<sup>3</sup>. Rheumatic heart disease in childhood, which is largely a result of poor access to health care and services, is a common cause of stroke among young adults in Southeast Asia and Africa<sup>4,5</sup>.

Although there is a well-established association between poverty and vulnerability to neurological disorders<sup>6</sup>, shockingly little is known regarding the pathophysiological processes that drive such links – even for exposures as ubiquitous as childhood malnutrition<sup>7,8</sup>. Given the substantial effect that childhood health and the environment (in the broadest sense) have on adult neurological health, longitudinal population-based studies of countries that have (and have not) met key United Nations' Millennium Development Goals are warranted. In the future, these experiments in poverty reduction over

<sup>1</sup>Epilepsy Division, Department of Neurology, University of Rochester, 265 Crittenden Boulevard, CU420694, Rochester, New York 14642-0694, USA. <sup>2</sup>Chikankata Epilepsy Care Team, Chikankata Hospital, Private Bag S2, Mazabuka, Zambia. <sup>3</sup>Department of Neurology, Yale University, P.O. Box 208018, New Haven, Connecticut, 06520-8018, USA. <sup>4</sup>Kenya Medical Research Institute, Box 614-40100, Kisumu, Kenya. <sup>5</sup>Department of Medicine, University of Ibadan, PMB 5016, Ibadan 200001, Nigeria. Correspondence should be addressed to G. B. e-mail: gretchen\_birbeck@urmc.rochester.edu.

relatively short time periods may also offer opportunities to formally evaluate well-developed theories regarding epigenetic responses to environmental change and the fetal origins of adult neurological conditions, particularly life-long cognitive impairment, and later onset dementia<sup>9</sup>.

Although malnutrition remains the principle nutritional challenge in child health in low-income countries, childhood obesity is reaching epic proportions in middle-income regions. Prevalence rates in South Africa (15%), Mexico (25%) and Brazil (23%), reflect the availability of relatively cheap, high-energy foods and increasingly sedentary lifestyles. The ongoing childhood obesity epidemic has resulted in unprecedented prevalence rates of hypertension among children and adolescents<sup>10</sup>. High body mass index, hypercholesterolemia and hypertension in childhood are associated with higher carotid artery intima-medial thickness in young adulthood – an established risk factor of cardiovascular disease and stroke<sup>11</sup>. Moreover, childhood hypertension is associated with hypertension as an adult<sup>12</sup>. Other adult cardiovascular risk factors include those that occur early in life with long-term consequences. Passive tobacco exposure in childhood causes lower high-density lipoprotein cholesterol<sup>13</sup> and childhood smoking is associated with greater carotid intima-medial thickness as an adult, even if the smoking habit ends before maturity<sup>14</sup>. Although exposure to established stroke risk factors from childhood undoubtedly increases an individual's overall risk of adult cerebrovascular disease, longitudinal studies are needed to quantify the impact of today's childhood obesity problem on the adult population of tomorrow.

## LATER LIFE DISEASE AND ADVANCES IN NEUROSCIENCE

Although risk factors for adult neurological disorders can occur in isolation in childhood, if problems such as obesity, tobacco use and hypertension begin in childhood or adolescence they often continue into adulthood. Cardiovascular risk factors and illiteracy play a major part in the growing burden of dementia in LMICs<sup>15</sup>. Without successful programmes aimed at poverty reduction and population-based interventions that curb harmful behaviours and encourage healthier lifestyles, the global prevalence of neurological disorders is likely to increase substantially.

The growing body of neurological research from LMICs has contributed substantially to, and at times transformed, our understanding of numerous neurological disorders, or the pathophysiological processes surrounding them. These contributions include observational<sup>16</sup> and epidemiological<sup>15</sup> studies, health-services research<sup>17</sup>, and clinical trials<sup>18</sup>, as well as laboratory-based sciences<sup>19</sup>. Many of these advances address neurological disorders that are common in both developed and less-developed settings<sup>20</sup>. Other contributions are focused on disorders more limited in geographical scope<sup>21</sup>, but these too offer advances on which future research can build.

## Advances in neuroscience basic sciences

The mysteries of neurodegeneration are being revealed through research collaborations between US investigators and LMIC scientists. Researchers from the University of Antioquia in Colombia have shown that viral vectors for RNA interference of cyclin-dependent kinase 5 can reduce neurofibrillary tangles in transgenic Alzheimer's disease mice<sup>22</sup>, and ongoing investigations are aimed at identifying human genes that modify the age at onset of Alzheimer's<sup>23</sup>. Researchers at New York University and the University of Uruguay are using salmonella-based A $\beta$  to develop a vaccine for Alzheimer's<sup>24</sup>. Argentinian investigators are also using recombinant adenoviral vectors and have shown that insulin-like growth factor (IGF) I gene therapy can restore hypothalamic dopaminergic function in senile rats<sup>25</sup>, and that IGF also improves motor function<sup>26</sup>. Collaborative studies often between researchers from multiple countries are contributing to our understanding of the genetic mechanisms of Parkinson's disease and stroke<sup>27</sup>. Further research of this nature is ongoing through the National Institutes of Health's H3Africa Initiative with a particular focus on stroke genomics in Africa.

**Epidemiological studies.** Epidemiological studies are well represented among those that contribute to our understanding of the global burden of nervous system disorders. Often initial studies offer more questions than answers, meaning subsequent investigations are required. For example, previous epidemiological estimates from developing regions persistently illustrated a 'gap' in epilepsy prevalence versus the lifetime prevalence rates, indicating that either people with epilepsy in developing countries have a higher remission rate or that they have higher rates of premature mortality than their counterparts in wealthier settings. As part of the International League against Epilepsy's China demonstration project, a prospective study of people with epilepsy in rural China demonstrated high rates of prime adult, seizure-related mortality despite readily accessible treatment<sup>28</sup>, indicating that this gap is from mortality and not disease remission. Multinational collaborations currently underway are elucidating the neurotoxicity of pesticides on both passively exposed children and adult agricultural workers<sup>29</sup>.

Although combined antiretroviral therapy is facilitating the transition of HIV from a fatal disease to a chronic condition, neurological disability remains a significant problem for people with HIV worldwide<sup>30</sup>. HIV-associated neurocognitive disorders (HANDs) are evident in almost half of those with HIV in the United States<sup>31</sup>. An international group working on the clade-specific effects of HIV on the central nervous system has gained important insights into the pathophysiology and prevention of HANDs. Studies in Africa and Latin America identified that HIV-associated neuropathy prevalence rates in those countries are substantially greater than in similar US populations, and point towards a central role for nutritional factors<sup>32</sup>. The studies clearly illustrate the need for further investigations aimed at directing future interventions. Interventions to optimize the cognitive outcomes for people with HIV in regions heavily affected by the disease will be needed to further limit the impact of HIV on human resources in LMICs.

**Implementation science.** Research across the life course is needed for neurological disorders, but translating the knowledge from research into action requires evidence-based approaches to implementation that are locally relevant. For example, researchers studying food-related neurotoxicity have successfully identified safer, feasible food preparatory measures, which have been taken up by communities<sup>33</sup>. Epilepsy is one of the commonest chronic neurological conditions in all settings regardless of economic status. The epilepsy community has called for improved screening and integrated care for the common psychiatric co-morbidities that often devastate people with epilepsy. Barriers to developmental and interventional studies that would facilitate this undertaking include time constraints, limited reimbursement and understaffing<sup>34</sup>. Epilepsy researchers in Zambia collaborating with faculty at the country's only psychiatric hospital have developed and validated a screening instrument for anxiety and depression that is feasible in their setting and appropriate for the people seeking care<sup>17</sup>. Prior to its implementation, less than 1% of people receiving epilepsy care were treated or referred for psychiatric care. After implementation, this number increased to 32% (ref. 17). Despite the recognition that the psychiatric and social morbidity of epilepsy contributes substantially to the burden of those affected, a recent systemic review found that few rigorous intervention studies aimed at decreasing this social morbidity have been conducted<sup>35</sup>. Local peer support groups in Zambia have been shown to decrease epilepsy-associated stigma and improve medication adherence among young people<sup>36</sup>. This work was accomplished in an environment with time, resource and staffing constraints. The approaches taken may offer guidance to researchers from higher income regions.

**Diagnostic advances.** Rigorous clinical investigations require the development or adaptation, and validation of instruments for categorizing and quantifying disease and disability. One of the chief barriers to research in LMICs is the availability of such instruments. In the past decade LMIC-based researchers have undertaken instrument devel-



opment for neuropsychiatric disorders and dementia in China; neuro-disabilities in Africa<sup>37</sup>; HIV-associated dementia and neuropathies in Kenya<sup>38</sup> and Zambia<sup>39</sup>; and stroke in Vietnam<sup>40</sup>, Argentina<sup>41</sup>, Portugal<sup>42</sup> and Nigeria<sup>43</sup>. Methodology and software have been developed to facilitate the use of neuroimaging to characterize and quantitatively describe novel conditions<sup>44</sup> that could prove important in this era of emerging infectious diseases.

**Clinical trials.** Although clinical trials conducted in LMICs might seem to be the study design that is least likely to inform care in high-income regions, a clinical trial addressing the use of invasive intracranial pressure monitoring conducted in Brazil and Ecuador has forced neurosurgeons and neurointensivists to reconsider long-held ideas of the value of such assessments<sup>45</sup>. Conducting clinical trials in settings that are less burdened by ingrained practice patterns and training conventions may offer crucial opportunities to collect the data that will allow us to examine whether common practices established in the absence of evidence are really warranted.

## OLDER ADULTS TO OLDEST OLD

Frailty and functional disability often afflict old age. Degenerative diseases that affect vision and hearing compound health problems in older people. But these are not inevitable characteristics of ageing. Neurological diseases of old age, for which substantial literature exists in LMICs, include stroke, dementia, Parkinson's disease and epilepsy. The quality of life of those affected is often poor and major depression may occur<sup>46</sup>.

Stroke is the most important cause of disability and a leading cause of death globally. A third of stroke patients in resource-limited settings die and another third are left with residual disability. The toll on economic productivity in these low-income countries that are poorly equipped to care for these individuals is considerable. According to the World Health Organization, stroke accounts for 55% of the disability-adjusted life years (DALYs) for neurological disorders, whereas Alzheimer's disease is responsible for 12.0% (ref. 47). Currently, stroke is the most important cause of hospital admission in many African countries and this has been adduced to the high frequency of untreated hypertension. Late presentation in hospitals, poor facilities for adequate care provision and prevention of complications contribute to the high mortality rate. The use of thrombolytic agents is limited because of exorbitant cost and late presentation in hospitals. Data on the risk factors for stroke have come from the INTERSTROKE study, which included participants from five countries in sub-Saharan Africa (Mozambique, Nigeria, South Africa, Sudan and Uganda). The frequency of intracerebral haemorrhage is relatively high in Africa compared with other regions; and the strongest association between hypertension and stroke was also reported in African participants (odds ratio, 4.96 (95% confidence interval, 3.11–7.91) versus 2.79 (95% confidence interval, 1.83–4.75) for developed countries. Factors associated with poor prognosis are impaired consciousness, swallowing problems, incontinence and chest infection based on data from Gambia<sup>48</sup>. Stroke will remain a problem unless urgent action is taken in health promotion and the provision of facilities for optimal care within the resources available. Stroke units can provide crucial support for meeting some of these challenges and are available in some tertiary care facilities<sup>49</sup>.

Dementia is the most common neurodegenerative disease that affects older adults and the oldest old. Data on disease burden is sparse in sub-Saharan Africa because of the relatively young population and the suspected concealment of cases within families<sup>50</sup>. The prevalence in African communities varies between 2.29% and 10.1%, although direct comparison of results between studies may be hampered by differences in the methodology used. The rates are higher in South American countries<sup>51</sup>, in East Asia, the Pacific and South Asia relative to the level of economic development<sup>52</sup>. According to Alzheimer's Disease International, 58% of those affected live in LMICs, where awareness is considered to be poor<sup>53</sup>. Alzheimer's is the predominant type of dementia, accounting for between 50% and 80% of cases<sup>54</sup>. Other types are vascular dementia, Lewy body dementia and frontotemporal lobar degeneration.

## BOX 1 | STUDIES CONDUCTED IN LMICS THAT INFORM RESEARCH IN HIGH-INCOME SETTINGS

- Studies of genetic risk factors for bipolar I disorder have the potential to improve our understanding of these factors in immigrant populations in the United States<sup>68</sup>.
- Proposed studies of stroke prevention in sickle cell disease in Africa<sup>69</sup> and studies of cognition and/or psychosis in familial populations<sup>70</sup> in Mexico can directly affect the health of Americans with these disorders.
- Proposed studies of the impact of HIV1 viral diversity on cognition and most studies evaluating the neurological and behavioural outcomes associated with environmental toxins or exposures would not be feasible in the United States, but the results of these studies provide crucial insights into health-related concerns directly applicable to US populations.
- Research on outcomes of guideline-based management of increased intracranial pressure (ICP) rather than invasive ICP monitoring<sup>71</sup> would not have been possible in the United States given the absence (appropriate or not) of equipoise among clinicians and entrenched practice patterns.
- Resource challenges facing research conducted in LMICs stimulate the development of more cost-effective diagnostics such as the tripolar concentric ring electrode based non-invasive transcutaneous focal stimulation for epilepsy<sup>72</sup>.

The risk factors for dementia are similar to those in western countries<sup>43</sup>. An intriguing observation was the lack of association between apolipoprotein E  $\epsilon 4$  allele and Alzheimer's<sup>54</sup> in sub-Saharan Africa. Furthermore, a recent publication from the Indianapolis-Ibadan group reported significantly lower rates of dementia in Africa compared with ethnically similar populations in the United States<sup>55</sup>. This finding could indicate a paradigm shift in dementia risk and could have a bearing on rising dementia burden as a consequence of changing dietary habits. Vascular factors have also been shown to increase dementia risk. Hypertension after the age of 65 increases the risk of incident dementia in elderly Yoruba Nigerians<sup>56</sup>. A novel presenilin 1 mutation was found to be associated with Alzheimer's in a South African family and was characterized by early-onset presentation<sup>57</sup>. Other types of dementia associated with neurodegenerative diseases, including frontotemporal dementia, probably exist in LMICs<sup>15</sup>.

Social isolation is linked with cognitive decline owing to limited cognitive stimulation. The extended family system that has so far provided a buffer for the care of older people in LMICs is being eroded by rural-urban migration and economic pursuits<sup>58</sup>. Furthermore, institutionalized care is frowned on in these countries because of the stigma of destitution. Carers face many problems, notably psychological stress and high rates of depression; carer distress, therefore, is an emerging problem. To obviate this, a model of care for LMICs is a creche-type service for those with dementia. Weight loss is another source of concern as this may be a pointer to the development of dementia<sup>59</sup>. Intervention studies in regions where institutionalized care for older people is not an established norm may offer insights for high-income settings.

Dementia is a public health priority and the use of medication is limited in LMICs; consequently, intensive efforts should be directed at identifying the risk factors and making interventions to stem the tide. Education and lifestyle changes (for example, the consumption of healthy diets, physical activity and engagement in community activities to reduce social isolation) can go a long way to preventing dementia. Mild cognitive impairment (MIC) is the intermediate stage between normal ageing and dementia. It is classified into amnesic and non-amnesic types, and each type can affect single or multiple cognitive domains<sup>60</sup>. Single-domain amnesic MCI is a precursor of Alzheimer's, and it is the most common. The prevalence of MCI varies between 10% and 40% in studies from Africa<sup>15,61–63</sup>. Vascular cognitive impairment follows stroke

with white matter lesions and medial temporal atrophy, and manifests as impairment of executive functions<sup>64</sup>. Important risk factors for MCI include age, hypertension, dyslipidaemia, illicit drug use and metabolic disorders<sup>15</sup>. HIV infection also causes mild neurocognitive dysfunction and should be investigated based on clinical judgement.

Parkinson's disease is the most common movement disorder that affects older people and has both motor and non-motor manifestations. Dementia and depression are common co-morbidities. In Guam, Parkinsonism is associated with Alzheimer's and amyotrophic lateral sclerosis. Exposure to neurotoxins such as cycasin seems to be important in disease aetiology; oxidative stress also plays a part<sup>65</sup>. Many neurotoxins have been implicated in tropical myelinopathies, including cyanogenic glycosides present in poorly fermented cassava<sup>66</sup> and  $\beta$ -N-oxalyl-L-alanine present in the chickling pea (*Lathyrus sativus*; also known as the grass pea), which if consumed during drought causes lathyrism<sup>67</sup>. These neurotoxins have not been implicated as risk factors for Parkinson's disease.

Symptomatic epilepsy in our global ageing population is redefining the epidemiology of a condition previously considered to be a disorder of children and young adults. In older people, the important aetiological factors are stroke, space occupying lesions, metabolic derangements, neurodegeneration and medication side effects. Neurocysticercosis and onchocerciasis are two tropical diseases that cause epilepsy in LMICs and should be looked for and treated. Prescribing cheap, cost-effective medications with minimal side effects and using simplified regimens could reduce the challenge posed by a treatment gap, which leaves most people with epilepsy in resource-limited settings untreated.

## INSIGHTS GAINED FROM WORK IN LMICS

This review of a decade of research into brain disorders in LMICs illustrates the rich diversity of disorders, geographical areas and populations studied. The studies used a variety of methods, approaches and study designs, ranging from qualitative research to clinical or epidemiological approaches, and to translational and basic science projects.

Although each study addresses issues of importance to brain disorders that are relevant to the host country, the body of research also has relevance to high-income populations. Global health research can inform our understanding of the health of all populations and improve health care in high-income settings by exploring the health of key underserved populations, such as recent immigrants or ethnic minorities; enabling research of disorders with high morbidity, but low prevalence, in higher income settings; empowering researchers to challenge existing paradigms; identifying approaches or interventions with better cost-effectiveness than existing standards of care; and identifying best practices with superior outcomes.

## OPPORTUNITIES TO BUILD ON RECENT ADVANCES

A decade of research has established a solid foundation on which to continue to build research collaborations and a body of knowledge about brain disorders in the developing world across the life course. Translating these scientific advances into individual or population-based prevention or treatment interventions that improve brain health is a crucial next step to achieve better health for individuals affected by brain disorders worldwide.

## FUTURE DIRECTIONS

Essentially all research programmes are conducted in a resource-limited setting because of finite budgets, human resources and research infrastructure. Every programme undertaken comes at the cost of the opportunity for another research programme, development activity or investment. As such, some principles for priority setting might be considered. Specific programme content and priorities may vary regionally and development of programmes should be undertaken through partnerships in which LMIC partners direct the content and purpose on the basis of local needs.

The impressive body of work emanating from LMICs and collaborations with researchers from these countries illustrates the value of such international partnerships to all involved (Box 1). Future research needs

**Table 1** | Research priorities to address the neurological burden of disease in low- and middle-income countries

Priorities	Example	Potential approach
Implementation studies of interventions or programmes found to have efficacy or be effective on a small scale	<ul style="list-style-type: none"> <li>Screening and treatment for depression and anxiety among people with, or mothers of those with, epilepsy</li> <li>Evaluate varying systems of dementia care for health, health-related quality of life and cost outcomes</li> </ul>	<ul style="list-style-type: none"> <li>Scale up of developed programmes with multifaceted, community-based assessments</li> <li>Randomized studies at a community level</li> </ul>
Intersection of neurological NCDs and infectious and/or MCH conditions that remain problematic in LMICs	<ul style="list-style-type: none"> <li>Stroke, epilepsy, cognitive impairment in people living with HIV/AIDS</li> <li>Life-course evaluation of the impact on adult health of childhood malnutrition and obesity</li> </ul>	<ul style="list-style-type: none"> <li>Accessing populations for study through established HIV services</li> <li>May require birth cohort, or a comparison of long-term outcomes for countries that did and did not meet Millennium Development Goals</li> </ul>
Maximize use of existing data	<ul style="list-style-type: none"> <li>Further elucidate the burden of neurological diseases within the framework needed for health-service delivery, treatment and secondary prevention</li> </ul>	<ul style="list-style-type: none"> <li>Re-framing of GBD 2010 data to quantify relative burden of neurological disorders as manifested rather than as viewed through the root cause Databases are available that are not focused on neurological health, but which capture relevant exposures and outcomes</li> </ul>
Study new disorders that may offer opportunities to gain understanding of mechanisms for neurological disease or injury that are more broadly relevant	<ul style="list-style-type: none"> <li>Nodding syndrome and konzo</li> </ul>	<ul style="list-style-type: none"> <li>Requires collaborations with clinicians, epidemiologists and basic scientists</li> </ul>
Clinical trials of potentially affordable interventions informed by insights gained in high-income setting and aimed at improving neurological outcomes	<ul style="list-style-type: none"> <li>Avoidance of hyperthermia after neurological injury informed by post-arrest and post-HIE cooling protocols</li> <li>Possible inclusion of aggressive antipyretics rather than overt cooling</li> </ul>	<ul style="list-style-type: none"> <li>Pragmatic clinical trials</li> </ul>

GBD, global burden of disease; HIE, hypoxic ischemia encephalopathy; LMICs, low- and middle-income countries; MCH, maternal and child health; NCDs, non-communicable diseases

to continue to address the shared burden of common conditions, including neurodegenerative disorders, stroke and epilepsy (Table 1). For neglected tropical diseases such as cysticercosis, rabies, trypanosomiasis and leprosy, methods developed for the study of rare and orphaned conditions (most of which are neurological) could be applied to these diseases. Although epidemiological and translational research across the spectrum of diseases and methodologies is still needed in resource-limited settings, an increased focus on interventional studies, such as clinical trials or population-based interventions is crucial. Interventions that are culturally relevant and feasible to implement in resource-limited settings, and which can be translated into effective policies that ultimately lead to improved health at both the individual and population levels are the essential next step and need to be prioritized to take full advantage of the body of knowledge generated so far.

- Hair, N. L., Hanson, J. L., Wolfe, B. L. & Pollak, S. D. Association of child poverty, brain development, and academic achievement. *J. Am. Med. Assoc. Pediatrics* **169**, 822–829 (2015).
- Wagner, R. G. et al. Prevalence and risk factors for active convulsive epilepsy in rural northeast South Africa. *Epilepsy Res.* **108**, 782–791 (2014).
- Whalley, L. J. et al. How the 1932 and 1947 mental surveys of Aberdeen schoolchildren provide a framework to explore the childhood origins of late onset disease and disability. *Maturitas* **69**, 365–372 (2011).
- Hashmi, M., Khan, M. & Wasay, M. Growing burden of stroke in Pakistan: a review of progress and limitations. *Int. J. Stroke* **8**, 575–581 (2013).
- Ntsekhe, M. & Damasceno, A. Recent advances in the epidemiology, outcome, and prevention of myocardial infarction and stroke in sub-Saharan Africa. *Heart* **99**, 1230–1235 (2013).
- Hamadani, J. D. et al. Cognitive deficit and poverty in the first 5 years of childhood in Bangladesh. *Pediatrics* **134**, e1001–e1008 (2014).
- Bergen, D. C. Effects of poverty on cognitive function: a hidden neurologic epidemic. *Neurology* **71**, 447–451 (2008).
- Bergen, D. C. & Silberg, D. Nervous system disorders: a global epidemic. *Arch. Neurol.* **59**, 1194–1196 (2002).

9. Whalley, L. J., Dick, F. D. & McNeill, G. A life-course approach to the aetiology of late-onset dementias. *Lancet Neurol.* **5**, 87–96 (2006).
10. Falkner, B. Hypertension in children and adolescents: epidemiology and natural history. *Pediatric Nephrol.* **25**, 1219–1224 (2010).
11. Juonala, M. et al. Influence of age on associations between childhood risk factors and carotid intima-media thickness in adulthood: the Cardiovascular Risk in Young Finns Study, the Childhood Determinants of Adult Health Study, the Bogalusa Heart Study, and the Muscatine Study for the International Childhood Cardiovascular Cohort (i3C) Consortium. *Circulation* **122**, 2514–2520 (2010).
12. LaRosa, C. & Meyers, K. Epidemiology of hypertension in children and adolescents. *Lebanese Med. J.* **58**, 132–136 (2010).
13. Raitakari, O. T. et al. Cardiovascular risk factors in childhood and carotid artery intima-media thickness in adulthood: the Cardiovascular Risk in Young Finns Study. *J. Am. Med. Assoc.* **290**, 2277–2783 (2003).
14. Napoli, C. et al. Influence of maternal hypercholesterolaemia during pregnancy on progression of early atherosclerotic lesions in childhood: Fate of Early Lesions in Children (FELIC) study. *Lancet* **354**, 1234–1241 (1999).
15. Kalara, R. N. et al. Alzheimer's disease and vascular dementia in developing countries: prevalence, management, and risk factors. *Lancet Neurol.* **7**, 812–826 (2008).
16. Nair, G. et al. Characterizing cognitive deficits and dementia in an aging urban population in India. *Int. J. Alzheimer's Dis.* **2012**, 673849 (2012).
17. Mbewe, E. K., Uys, L. R. & Birbeck, G. L. The impact of a short depression and anxiety screening tool in epilepsy care in primary health care settings in Zambia. *Am. J. Trop. Med. Hygiene* **89**, 873–874 (2013).
18. Joray, M. L. et al. Zinc supplementation reduced DNA breaks in Ethiopian women. *Nut. Res.* **35**, 49–55 (2015).
19. Castro-Alvarez, J. F., Uribe-Arias, S. A., Kosik, K. S. & Cardona-Gomez, G. P. Long- and short-term CDK5 knockdown prevents spatial memory dysfunction and tau pathology of triple transgenic Alzheimer's mice. *Front. Aging Neurosci.* **6**, 243 (2014).
20. Van Naarden Braun, K. et al. Evaluation of a methodology for a collaborative multiple source surveillance network for autism spectrum disorders — Autism and Developmental Disabilities Monitoring Network, 14 sites, United States, 2002. *MMWR Surveillance Sum.* **56**, 29–40 (2007).
21. Nishioka, K. et al. Glucocerebrosidase mutations are not a common risk factor for Parkinson disease in North Africa. *Neurosci. Lett.* **477**, 57–60 (2010).
22. Piedrahita, D. et al. Silencing of CDK5 reduces neurofibrillary tangles in transgenic Alzheimer's mice. *J. Neurosci.* **30**, 13966–13976 (2010).
23. Lalli, M. A. et al. Exploratory data from complete genomes of familial Alzheimer disease age-at-onset outliers. *Human Mut.* **33**, 1630–1634 (2012).
24. Boutajangout, A. et al. Diminished amyloid-beta burden in Tg2576 mice following a prophylactic oral immunization with a salmonella-based amyloid-beta derivative vaccine. *J. Alzheimer's Dis.* **18**, 961–972 (2009).
25. Herenu, C. B. et al. Restorative effect of insulin-like growth factor-I gene therapy in the hypothalamus of senile rats with dopaminergic dysfunction. *Gene Ther.* **14**, 237–245 (2007).
26. Nishida, F. et al. Restorative effect of intracerebroventricular insulin-like growth factor-I gene therapy on motor performance in aging rats. *Neuroscience* **177**, 195–206 (2011).
27. Atadzhanov, M. et al. Association of the APOE, MTHFR and ACE genes polymorphisms and stroke in Zambian patients. *Neurology Int.* **5**, e20 (2013).
28. Ding, D. et al. Premature mortality in people with epilepsy in rural China: a prospective study. *Lancet Neurol.* **5**, 823–827 (2006).
29. London, L. et al. Neurobehavioral and neurodevelopmental effects of pesticide exposures. *Neurotoxicology* **33**, 887–896 (2012).
30. Rumbaugh, J. A., Steiner, J., Sacktor, N. & Nath, A. Developing neuroprotective strategies for treatment of HIV-associated neurocognitive dysfunction. *Future HIV Ther.* **2**, 271–280 (2008).
31. Heaton, R. K. et al. HIV-associated neurocognitive disorders persist in the era of potent antiretroviral therapy: CHARTER Study. *Neurology* **75**, 2087–2096 (2010).
32. Birbeck, G. L. et al. Neuropsychiatric and socioeconomic status impact antiretroviral adherence and mortality in rural Zambia. *Am. J. Trop. Med. Hyg.* **85**, 782–789 (2011).
33. Banea, J. P. et al. Effectiveness of wetting method for control of konzo and reduction of cyanide poisoning by removal of cyanogens from cassava flour. *Food Nut. Bull.* **35**, 28–32 (2014).
34. Asato, M. R., Caplan, R. & Hermann, B. P. Epilepsy and comorbidities — what are we waiting for? *Epilepsy Behav.* **31**, 127–128 (2014).
35. Fiest, K. M., Birbeck, G. L., Jacoby, A. & Jette, N. Stigma in epilepsy. *Current Neurol. Neurosci. Rep.* **14**, 444 (2014).
36. Elafros, M. A. et al. Peer support groups as an intervention to decrease epilepsy-associated stigma. *Epilepsy Behav.* **27**, 188–192 (2013).
37. Bower, J. H. et al. Validity of a screening instrument for neurologic disability in resource-poor African communities. *J. Neurolog. Sci.* **320**, 52–55 (2012).
38. Kwasa, J. et al. Lessons learned developing a diagnostic tool for HIV-associated dementia feasible to implement in resource-limited settings: pilot testing in Kenya. *PLoS ONE* **7**, e32898 (2012).
39. Kvalsund, M., Chidumayo, T., Hamel, J., Heimbürger, D. & Birbeck, G. L. Prevalence and comorbid factors associated with distal symmetric polyneuropathies in HIV+ and HIV- adults in urban and rural Zambia. *Neurology* **84**, P5.049 (2015).
40. Tirschwell, D. L. et al. A prospective cohort study of stroke characteristics, care, and mortality in a hospital stroke registry in Vietnam. *BMC Neurology* **12**, 150 (2012).
41. Sposato, L. A. et al. Program for the epidemiological evaluation of stroke in Tandil, Argentina (PREVISTA) study: rationale and design. *Int. J. Stroke* **8**, 591–597 (2013).
42. Ferreiro, K. N., Santos, R. L. & Conforto, A. B. Psychometric properties of the Portuguese version of the Jebsen-Taylor test for adults with mild hemiparesis. *Revista Brasileira Fisioterapia* **14**, 377–382 (2010).
43. Ochay, B. & Thacher, T. D. Risk factors for dementia in central Nigeria. *Aging Mental Health* **10**, 616–620 (2006).
44. Potchen, M. J. et al. NeuroInterp: a method for facilitating neuroimaging research on cerebral malaria. *Neurology* **81**, 585–588 (2013).
45. Chesnut, R. M. et al. A trial of intracranial-pressure monitoring in traumatic brain injury. *N. Engl. J. Med.* **367**, 2471–2481 (2012).
46. Gureje, O., Ogunniyi, A., Kola, L. & Afolabi, E. Functional disability in elderly Nigerians: Results from the Ibadan Study of Aging. *J. Am. Geriatrics Soc.* **54**, 1784–1789 (2006).
47. World Health Organization. *Neurologic Disorders: Public Health Challenges*. (WHO, 2006).
48. Walker, R. W., Rolfe, M., Kelly, P. J., George, M. O., James, O. F. Mortality and recovery after stroke in the Gambia. *Stroke* **34**, 1604–1609 (2003).
49. Aiwansoba, I. F. & Chukwuyem, O. W. Early post-acute stroke seizures: clinical profile and outcome in a Nigerian stroke unit. *Ann. African Med.* **13**, 11–15 (2014).
50. Ineichen, B. The epidemiology of dementia in Africa: a review. *Social Sci. Med.* **50**, 1673–1677 (2000).
51. Llibre Rodriguez, J. J. et al. Prevalence of dementia in Latin America, India, and China: a population-based cross-sectional survey. *Lancet* **372**, 464–474 (2008).
52. Chandra, V. et al. In *Disease Control Priorities in Developing Countries*. 2nd edn (eds Jamison, D. T. et al.) (World Bank, 2006).
53. Alzheimer's Disease International. *The Global Economic Impact of Dementia* (ADI, 2010).
54. Gureje, O. et al. APOE e4 is not associated with Alzheimer's disease in elderly Nigerians. *Ann. Neurol.* **59**, 182–185 (2006).
55. Hendrie, H. C. et al. Prevalence of Alzheimer's disease and dementia in two communities: Nigerian Africans and African Americans. *Am. J. Psych.* **152**, 1485–1492 (1995).
56. Ogunniyi, A. et al. Hypertension and incident dementia in community-dwelling elderly Yoruba Nigerians. *Acta Neurolog. Scand.* **124**, 396–402 (2011).
57. Heckmann, J. M. et al. Novel presenilin 1 mutation with profound neurofibrillary pathology in an indigenous Southern African family with early-onset Alzheimer's disease. *Brain* **127**, 133–142 (2004).
58. Ogunniyi, A. et al. Weight loss and incident dementia in elderly Yoruba Nigerians: a 10-year follow-up study. *Int. Psychogeriatrics* **23**, 387–394 (2011).
59. Ogunniyi, A. et al. Caring for individuals with dementia: the Nigerian experience. *West African J. Med.* **24**, 259–262 (2005).
60. Petersen, R. C. Mild cognitive impairment as a diagnostic entity. *J. Int. Med.* **256**, 183–194 (2004).
61. Guerchet, M. et al. Cognitive impairment and dementia in elderly people living in rural Benin, west Africa. *Dementia Geriatric Cognitive Dis.* **27**, 34–41 (2009).
62. Baiyewu, O. et al. Cognitive impairment in community-dwelling older Nigerians: clinical correlates and stability of diagnosis. *Eur. J. Neurol.* **9**, 573–580 (2002).
63. Coume, M. et al. Estimate of the prevalence of cognitive impairment in an elderly population of the health center of Senegalese national retirement institution. *Geriatric Psychologie Neuropsychiatrie Vieillesse* **10**, 39–46 (2012).
64. Akinyemi, R. O. et al. Profile and determinants of vascular cognitive impairment in African stroke survivors: the CogFAST Nigeria Study. *J. Neurol. Sci.* **346**, 241–249 (2014).
65. Spencer, P. S. Food toxins, ampa receptors, and motor neuron diseases. *Drug Metabol. Rev.* **31**, 561–587 (1999).
66. Tshala-Katumbay, D. D. & Spencer, P. S. Toxic disorders of the upper motor neuron system. *Handbook Clin. Neurol.* **82**, 353–372 (2007).
67. Getahun, H., Mekonnen, A., Teklehaimanot, R. & Lambein, F. Epidemic of neuroleptism in Ethiopia. *Lancet* **354**, 306–307 (1999).
68. Mansour, H. et al. Consanguinity associated with increased risk for bipolar I disorder in Egypt. *Am. J. Med. Genet. B* **150B**, 879–885 (2009).
69. Galadanci, N. A. et al. Primary stroke prevention in Nigerian children with sickle cell disease (SPIN): challenges of conducting a feasibility trial. *Pediat. Blood Cancer* **62**, 395–401 (2015).
70. Ringman, J. M. What the study of persons at risk for familial Alzheimer's disease can tell us about the earliest stages of the disorder: a review. *J. Geriatric Psych. Neurol.* **18**, 228–233 (2005).
71. Rubiano, A., Puyana, J., Mock, C., Bullock, M. & Adelson, P. Strengthening neurotrauma care systems in low and middle income countries. *Brain Injury* **27**, 262–272 (2013).
72. Makeyev, O., Luna-Munguia, H., Rogel-Salazar, G., Liu, X. & Besio, W. G. Noninvasive transcranial focal stimulation via tripolar concentric ring electrodes lessens behavioral seizure activity of recurrent pentylenetetrazole administrations in rats. *IEEE Trans Neural Systems Rehabilitation Engineering* **21**, 383–390 (2013).

#### ACKNOWLEDGEMENTS

This work emanated from the Tenth Anniversary BRAIN Disorders in the Developing World Symposium. We are grateful to all participants and to National Institutes of Health and the Fogarty International Center for sponsoring the meeting.

#### COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests. Financial support for publication has been provided by the Fogarty International Center.

#### ADDITIONAL INFORMATION



This work is licensed under the Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0>



## REVIEW OPEN

# Global research challenges and opportunities for mental health and substance-use disorders

Florence Baingana<sup>1</sup>, Mustafa al'Absi<sup>2</sup>, Anne E. Becker<sup>3</sup> & Beverly Pringle<sup>4</sup>

The research agenda for global mental health and substance-use disorders has been largely driven by the exigencies of high health burdens and associated unmet needs in low- and middle-income countries. Implementation research focused on context-driven adaptation and innovation in service delivery has begun to yield promising results that are improving the quality of, and access to, care in low-resource settings. Importantly, these efforts have also resulted in the development and augmentation of local, in-country research capacities. Given the complex interplay between mental health and substance-use disorders, medical conditions, and biological and social vulnerabilities, a revitalized research agenda must encompass both local variation and global commonalities in the impact of adversities, multi-morbidities and their consequences across the life course. We recommend priorities for research — as well as guiding principles for context-driven, intersectoral, integrative approaches — that will advance knowledge and answer the most pressing local and global mental health questions and needs, while also promoting a health equity agenda and extending the quality, reach and impact of scientific enquiry.

*Nature* 527, S172–S177 (19 November 2015), DOI: 10.1038/nature16032

This article has not been written or reviewed by *Nature* editors. *Nature* accepts no responsibility for the accuracy of the information provided.

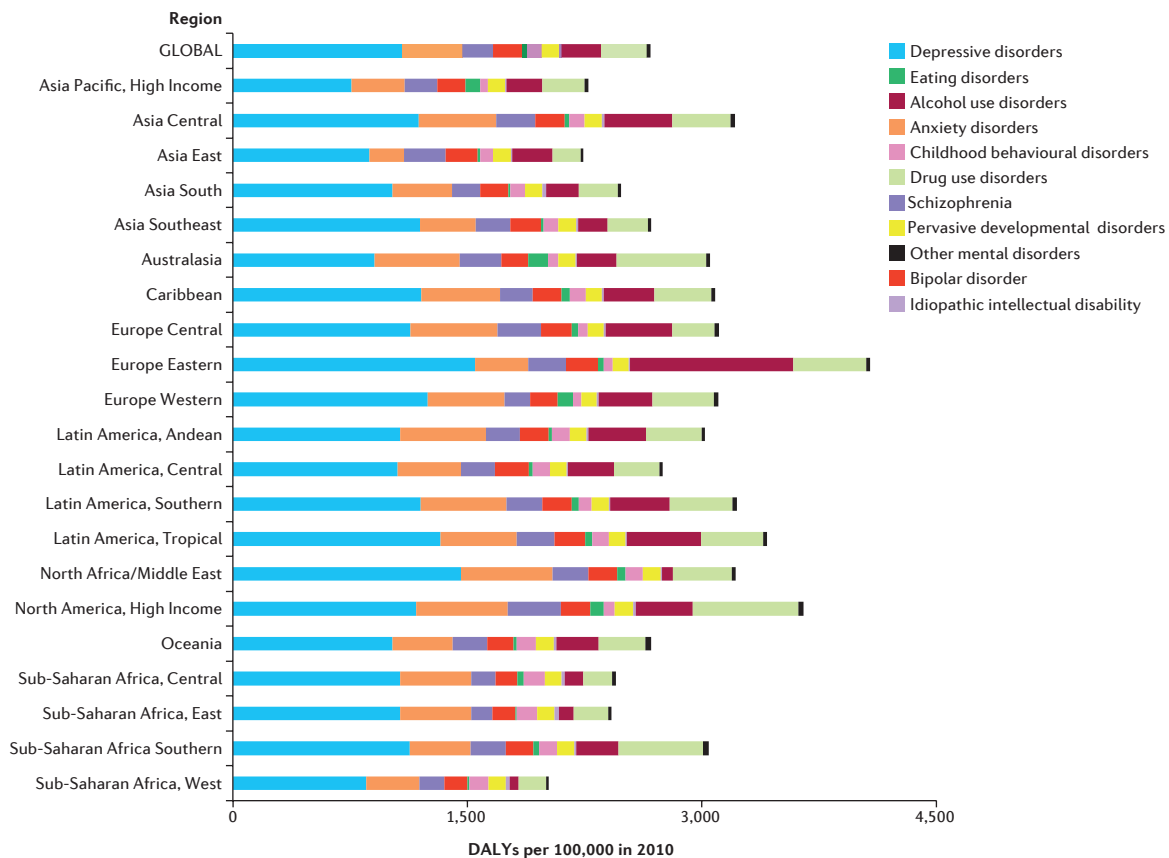
The global mental health landscape has transformed over the past 25 years because of the higher visibility of the burden of mental health and substance-use disorders<sup>1</sup>. These disorders comprise 7.4% of global disability-adjusted life years (DALYs) and 22.7% of global years lived with disability (YLDs)<sup>2</sup> (Supplementary Information)<sup>3</sup>. The main contributors worldwide are depression and dysthymia (9.6% of all YLDs); anxiety (3.5% of all YLDs); and schizophrenia, substance-use disorders and bipolar disorder (just over 2% of all YLDs). Alcohol and substance-use disorders come in second for most of the developing world, more so for southern Africa (drug use) and Eastern Europe (alcohol)<sup>2</sup>. The burden of mental health and substance-use disorders is predicted to increase worldwide in coming decades, and the steepest rise can be expected in low- and middle-income countries (LMICs) as a result of rising life expectancy, population growth and under-resourced health care<sup>4</sup>. For example, simulation models predict a 130% increase in associated health burden of alcohol and substance misuse in sub-Saharan Africa by 2050 as a result of population growth and ageing<sup>5</sup>. As substantial as they are, conventional health metrics do not capture additional social burdens attached to living with mental illness. Untreated mental health disorders are associated with a high economic burden<sup>6</sup>. Furthermore, pervasive stigma and human rights violations compound the suffering associated with these disorders and exacerbate social vulnerabilities<sup>7–9</sup>.

As the health, social, economic and human costs of mental and substance-use disorders become increasingly better documented, political will and multilateral commitments to scale up mental health care in LMICs have grown. The World Health Organization has introduced a series of policy initiatives that articulate both high-level aspi-

ration and pragmatic guidance for mental health and substance-use services delivery in LMICs. The most recent, the Global Mental Health Action Plan 2013–2020, challenges member states, partners and the Secretariat to collectively meet ambitious goals by the year 2020, including increasing mental health care coverage by 20% for severe mental health illness and reducing national suicide rates by 10% (ref. 10). A consideration of the timeline of these landmark events — including the roll out of a number of key funding and policy initiatives that target the persisting resource gaps — illustrates the substantial momentum in integrating mental health into the broader global health agenda that has occurred over the past few decades (See Supplementary Information).

An interactive map, depicting the broad geographical distribution of current, promising initiatives in global mental health is available at <http://www.nimh.nih.gov/responsive/map.shtml>. Policy and programmatic initiatives have laid a foundation for strengthened global mental health services by developing an initial consensus scientific agenda that focuses energies and funding on the most crucial research for building an empirical base. Key funding initiatives have supported research to leverage scarce resources and improve access through task sharing, integration of mental health care into existing primary health-care infrastructure and enhancement of diagnostic assessment. Increasingly, resources have been allocated to build in-country research capacities and strengthen collaboration through institutional partnerships<sup>11</sup>. Complementary graduate-level training programmes in global mental health have also emerged, although mental health specialty training, as a track, remains relatively under-represented among other global health domains<sup>12,13</sup>.

<sup>1</sup>Makerere University School of Public Health, PO Box 7072, Kampala, Uganda. <sup>2</sup>Duluth Medical Research Institute (DMRI), University of Minnesota Medical School, 311-1035 University Drive, Duluth, Minnesota 55812, USA. <sup>3</sup>Department of Global Health and Social Medicine, Harvard Medical School, 641 Huntington Avenue, Boston, Massachusetts 03115, USA. <sup>4</sup>Office for Research on Disparities & Global Mental Health, National Institute of Mental Health, 6001 Executive Boulevard, Room 7207, Bethesda, Maryland 20892, USA. Correspondence should be addressed to F.B. e-mail: kamayonza@gmail.com.



**Figure 1** | Burden of risk associated with substance-use disorders varies across the world as disability adjusted life years (DALYs). Reprinted with permission from ref.4.

### Key research gaps and challenges

The global health burden of mental health disorders is exacerbated by the growing concurrent problems associated with substance misuse. Substance use and exposure to addictive drugs have chronic and profound effects on neurobehavioural and neurodevelopmental functions. In LMICs, the socioecology of poverty, malnutrition, political conflicts and poor health systems influence the epidemiology, as well as the adverse outcomes, that result from substance misuse. Additional challenges associated with co-morbidity stem from its augmentation of clinical burden, through increased risk for relapse, other infectious and medical complications, and economic hardship and homelessness. In this context, co-morbid substance use and mental illnesses in particular may contribute to increasing health burden. The prevalence of substance-use disorders has escalated in recent decades, reaching 5.4% of the total disease burden and 9.1% when tobacco use is included<sup>14</sup>.

Individuals with substance-use disorders are also likely to have other mental health problems, including depression and schizophrenia<sup>4,15</sup>. Similarly, a large proportion of people with mental illnesses also have substance-use disorders<sup>16,17</sup>. Research that investigates the relationship between mental illness and substance-use disorder has yielded mixed findings, with some support for causal relationships in both directions as well as for shared genetic, environmental, social and cultural risk factors. For example, cannabis use is linked to a risk of developing psychotic illness<sup>18</sup>. Conversely, mental illness may increase the risk of substance misuse; individuals may 'self-medicate' with alcohol, tobacco or amphetamines as a means of coping with distress and negative affects<sup>19,20</sup>. Some factors, including genetic vulnerabilities, traumatic exposures and stress, may confer risk for both conditions<sup>21,22</sup>. Diagnosis and treatment of co-morbid substance misuse and mental health illness remains a significant challenge, particularly in LMICs. The burden of this co-morbidity is further exacerbated by the

increased clinical complexity that stems from resistance to treatment, risk of relapse, vulnerability to other infectious and medical complications, and increased economic hardship and homelessness.

The burden and configuration of risk associated with substance-use disorders and co-morbid mental illness seem to vary across the world (Fig. 1)<sup>14</sup>. Although alcohol and opioid problems are escalating in Europe, Africa and Asia, problems associated with amphetamines and cannabis are more prevalent in Asia, North America and Europe. Cocaine use is prevalent in North America and Europe, whereas misuse of indigenous psychoactive substances is prevalent in other regions, such as the use of khat in parts of Africa and the Middle East and that of coca leaves in South America<sup>23</sup>. Notably, existing knowledge gaps may underestimate the impact of substance-use disorders<sup>24</sup>. The full extent of adverse mental health and social impacts of substance-use disorders such as alcohol use during pregnancy and fetal alcohol spectrum disorders<sup>25</sup> remain incompletely understood.

Mental health and substance-use disorders also frequently co-occur with other diseases, increasing associated morbidity and mortality risk<sup>26,27</sup>. It is not uncommon for individuals with HIV/AIDS or non-communicable diseases such as hypertension, diabetes and cardiovascular disease to also have symptoms of depression or anxiety and to use alcohol or other drugs to excess. Attention deficit hyperactivity disorder has been associated with risky sexual behaviours that can result in transmission of HIV/AIDS. These interdependent illnesses stem from common risk factors, such as childhood adversity; and bidirectional influences, such as poor treatment adherence<sup>28</sup> and increased engagement in risky behaviour<sup>29–31</sup>. Growing awareness of the complex interplay between mental illness and the increasing burden of chronic disease globally has prompted research that examines the effects of depression on adherence to medical treatments and the effects of integrated care — co-treatment of high blood pressure and depression, for example — on the outcomes of both of the co-occurring illnesses (see

for example refs 32–34). A life-course approach to risk reduction that takes into account risks that occur in childhood and early adulthood, and that promotes a healthy lifestyle, and early recognition and treatment of mental and substance-use disorders is essential to curtail the long-term negative impacts of many preventable health risks.

## TREATMENT GAP

The proportion of people who need, but do not receive care is especially high in LMICs<sup>35,36</sup>. The inadequate resourcing of mental health care in LMICs has been widely documented and critiqued. For example, on average less than 3% of public health resources are allocated to specific mental health care in LMICs, with even less (around 1%) in Africa and Asia<sup>37</sup>. Most LMICs have far fewer health-care professionals than they need to deliver mental health and substance-use interventions to everyone who needs them<sup>38,39</sup>. Scaling up services will require more than training additional psychiatrists, psychologists and psychiatric nurses, however, strategic leveraging of scarce resources will also be necessary. In particular, task shifting — delegating health-care tasks from specialists to various non-specialist health professionals and other health workers — has shown promise for certain mental health and substance-use interventions<sup>40–43</sup>. In addition, the integration of mental health services into primary health-care delivery settings through community-based and task-sharing approaches can both help to reduce burden on carers and improve access and the coordination of care. Mental health services and health-system strengthening, and in particular, task shifting, as well as organization and ways of delivering community-based mental health services in LMICs should be prioritized for research.

There is also a substantial gap in scientific knowledge for preventing and treating mental health and substance-use disorders. In addition, what is currently known is often not applicable to low-resource regions. Intervention strategies to address substance-use disorders have improved over recent decades, but have had limited success in achieving total recovery and have limited coverage in LMICs<sup>15</sup>. Moreover, resources for providing these interventions are constrained or lacking in most LMICs<sup>15,24</sup>. Models for improving availability and access to effective mental health care emphasize the integration of both prevention and treatment services within primary care systems. This has been a core approach taken by the WHO Mental Health Gap Action Program (mhGAP)<sup>44,45</sup>.

Most published clinical trial data on therapeutics for mental health disorders are based on research conducted in high-income countries<sup>46–48</sup>. In the absence of region-specific empirical data, deployment of these therapeutic strategies in LMICs is a reasonable pragmatic compromise in the short term when informed by local knowledge, and pending rigorous and systematic evaluation. Local research on clinical effectiveness of these treatments and implementation research on how to deliver these therapies and scale them up are urgently needed.

## Priorities for advancing the global mental health agenda

Our recommendations build on the strong base of empirical evidence and previous consensus statements and reports that have articulated principles, needs and priorities that should inform a robust research agenda (Table 1). The predominant focus of global mental health research is currently on health services and implementation research, areas that align well with efforts to close treatment gaps and that must continue to be strengthened. Whereas we regard these contributions as formative and arguably the most pragmatic and exigent in the short term, they should not pre-empt a more ambitious scope of scientific inquiry that ranges from basic sciences to health policy. Innovation should encompass much more than strategies to leverage scarce resources lest the scope of progress in the field be consigned to improving the efficiency of old models of care delivery. Instead, complementary and parallel lines of context-driven research should also aim to advance the scientific understanding of aetiology, population-specific phenotypic variation in presentation and course, and differential

response to therapeutics through promising avenues in neuroscience, biomarkers, genetics and epigenomics.

## EPIDEMIOLOGY

Epidemiological research is crucial to better understand the differential risk factors and burden of mental health and substance-use disorders across diverse geographical regions and social contexts. Refinement of approaches to diagnostic assessment that are both locally valid and relatable to global classification is essential to more effective and efficient case identification, particularly in the hands of non-specialists. Such advances will generate more accurate estimates of health burdens and salient risk factors on which local health policymakers can draw. In addition, research is needed to better define the health and social impacts of syndemic mental health disorders, substance-use disorders and medical diseases, as well as to understand how social adversities moderate and mediate risk of onset, severity and course. Such research will inform optimal strategies for prevention, treatment and follow-up care for individuals with these co-morbidities.

## BASIC SCIENCE RESEARCH

The research agenda to address the global burden of mental health and substance-use disorders should build on recent advances in the field of basic neuroscience, biomarkers, proteomics, and genetics and epigenetics. For example, research in the past decade has identified molecular and structural markers connected with mental health and substance-use disorders<sup>49</sup>. These include protein alterations in the form of upregulation of a 40-amino acid VGF-derived peptide and the downregulation of apoA1 protein in schizophrenia<sup>50</sup>. Hormonal and physiological alterations in stress- and appetite-related neuropeptides have also been pursued in the context of addiction and treatment outcome<sup>51–53</sup>. There has also been significant interest in epigenomics and how it could advance our understanding and use of biomarkers. Epigenomic modifications affect gene expression, and involve multiple molecular steps, including DNA methylation<sup>54</sup>. In light of evidence that indicates a role for epigenetic mechanisms in modifying genes that increase propensity for drug use and mental illness, it is important to develop a means by which this approach could be harnessed to improve the validity and reliability of diagnostic measures as well as to help to tailor interventions to the individual. Research that considers the diversity of environmental exposures and gene-environment interactions across different settings can advance the utility of these markers to confirm diagnosis and to predict treatment outcome. Furthermore, such markers may be useful in identifying those at high risk so that measures can be applied to prevent initial risk or onset, or slow down or prevent progression towards psychopathology. The use of such approaches should also parallel the development of conceptual models to guide understanding of the complex, multidimensional aetiology of mental health and substance-use disorders. To that end, global research that focuses on mental health and substance-use disorders should take into account how genetics and exposure to environmental toxins interact with social, cultural and environmental conditions to moderate the risk of these disorders.

## HEALTH DELIVERY AND IMPLEMENTATION RESEARCH

Four out of the top five research priorities identified in the grand challenges statement — developed by a consortium of researchers, advocates and clinicians with funding from the US National Institute of Mental Health (NIMH) and the Global Alliance for Chronic Diseases — fall in the domain of enhancing the quality of, and access to, mental health care<sup>55</sup>. This call to invest in health services and implementation research is in response to identified treatment gaps as well as their numerous antecedents, such as weak health systems, shortfalls in human and financial resources, and social structural barriers to care. There is ample evidence for science-based care and the integration of mental health services into primary health care. However, we still lack crucial knowledge on how best to disseminate and implement evidence-based



**Table 1** | Summary of priority research areas and guiding principles for approaches to reduce the global burdens of mental and substance-use disorders.

Research domain	Priority areas
<b>Epidemiology</b>	<ul style="list-style-type: none"> <li>Refine approaches to diagnostic assessment across diverse local contexts by enhancing their local validity within universal frameworks, and promoting and evaluating their clinical utility for non-specialist health professionals and other health workers, including in a wide variety of health-care and community-based settings</li> <li>Improve accuracy of estimates of health and related social and economic burdens</li> <li>Advance understanding of local contextual factors that mediate and moderate risk, course of illness, and recovery for mental and substance-use disorders by identifying synergies among co-morbid mental health and substance-use disorders, and medical conditions relating to risk, illness trajectories and treatment outcomes; understanding context-specific impacts of social adversities and local resources that promote resilience; illuminating social vulnerabilities that impede access to care; examining and comparing population-specific phenotypic variants of major mental health and substance-use disorders to enhance understanding of genetic, developmental, environmental, social and cultural contributions to aetiology, course and presentation</li> <li>Evaluate health, social and economic benefits of general health, mental health and substance-use disorder therapeutic intervention, prospectively across the life course and the next generation</li> </ul>
<b>Basic sciences</b>	<ul style="list-style-type: none"> <li>Advance understanding of gene–environment interactions with the scientific benefits that variation across diverse contexts provides</li> <li>Identify moderators of course of illness, therapeutic response, remission, relapse and recovery</li> </ul>
<b>Health-care delivery and implementation</b>	<ul style="list-style-type: none"> <li>Optimize effectiveness of task-sharing models for community-based case-finding and treatment delivery</li> <li>Understand and resolve barriers to integration of mental health and substance-use disorder assessment and treatment in primary care settings and in non-health platforms (for example, schools, housing and the criminal justice system)</li> <li>Mitigate barriers to care access</li> <li>Develop and improve coordinated approaches to strategic preventive interventions, monitoring and targeted treatments over the life course and across disorders</li> <li>Evaluate successful delivery models at scale and their adaptation to diverse local contexts</li> <li>Advance understanding of factors that mediate and moderate successful care delivery at scale</li> </ul>
<b>Translational research and health policy</b>	<ul style="list-style-type: none"> <li>Illuminate 'sideways' impacts of economic, housing, criminal justice and education policies on mental health and substance-use disorders</li> <li>Develop evidence-based strategies that accelerate the uptake of mental health research findings by policymakers</li> </ul>
Guiding principles	Rationale for approach
<b>Responsive to exigent needs related to burden of disease and associated treatment gaps</b>	<ul style="list-style-type: none"> <li>Consistent with the broad global health agenda of reducing health burdens, research recommendations focus on a globalizing framework and leveraging strategy that can bring effective interventions to a larger scale</li> <li>This pragmatic focus should not eclipse the complementary relevance and importance of extending research into comparative differences in mental and substance-use disorders across diverse populations and social contexts</li> </ul>
<b>Context driven</b>	<ul style="list-style-type: none"> <li>Diversity across local social, economic, environmental, cultural and political contexts of health and health care has profound implications for risk, resilience, presentation, consumer demand, therapeutic outcomes, and impacts of mental and substance-use disorders; however, these factors remain incompletely understood</li> <li>Engagement of local knowledge is an essential complement to universalizing frameworks and the aggregate global empirical base in developing a fully contextualized understanding and optimization of local responses to mental health and substance-use disorders as well as in formulating and pursuing the most salient local research priorities</li> </ul>
<b>Integrative across primary and specialty clinical care domains and the life course</b>	<ul style="list-style-type: none"> <li>Mental disorders, substance-use disorders and medical conditions are frequently co-morbid and thus development, evaluation and implementation of delivery models that enhance coordination and integration of care — across disorders, conditions and the life course — are likely to reduce the burden and impacts of these disorders</li> </ul>
<b>Inter-sectoral</b>	<ul style="list-style-type: none"> <li>Improved links between scientific evidence and policymaking can enable and accelerate translation from research to practice</li> <li>Cross-sector planning can promote efficiencies and achievement of health sector goals by anticipating necessary preliminary or parallel actions in finance, education and other sectors</li> </ul>
<b>Capacity building</b>	<ul style="list-style-type: none"> <li>Research activities should encompass an agenda and plan for in-country capacity building to grow the global research work force, extend the quality and reach of scientific enquiry, and promote a health equity agenda</li> <li>Capacity-building should be bidirectional and collaborative, when possible, to tap complementary expertise and experience in research, clinical practice, implementation, policy and other relevant dimensions that will advance the local and global mental health agenda</li> </ul>

mental health interventions in resource-poor contexts, including those characterized by the extreme social adversities associated with political conflict, displacement and destitution. Future research is therefore necessary to rigorously evaluate and optimize effectiveness of task sharing, integration of mental health into primary care, and deployment of the mhGAP algorithms at larger scale and across diverse social settings<sup>41,56</sup>.

Key strategies for expanding access to high-quality mental health care in LMICs come from models that are successful in leveraging scarce resources in other clinical domains. However, challenges unique to care delivery for mental health and substance-use disorders warrant special attention and innovation. These include how to improve diagnostic assessment and population health surveillance, given the heterogeneous and sometimes opaque presentations of signs and symptoms across diverse social and cultural contexts<sup>57,58</sup>; how to address the social and cultural factors, especially stigma, that hinder access to care and may prevent patients with mental illnesses and substance-use problems from using the resources available for prevention and treatment; how to mitigate social vulnerabilities, such as poverty and gender-based violence that elevate risk of mental disorders, while building on sociocultural resources that promote coping and resilience; how to develop coordinated approaches to strategic preventive interventions, monitoring and targeted treatments over the life course and across disorders, given developmental trajectories of mental health and substance-use disorders, and their harmful symbiosis with other chronic conditions and vulnerabilities; and how to rapidly scale up effective interventions to close the treatment gap in resource constrained

environments<sup>59–61</sup>. Priorities for global mental health research resonate with the global health agenda, with its focus on reducing health burdens<sup>62</sup>. In this respect, a globalizing framework aimed at developing approaches that are effective when scaled up and implemented across geographically and socially diverse settings and populations reflect pragmatic goals of responding to pressing needs. We emphasize, however, that closing the prevailing treatment gaps for mental health and substance-use disorders will also depend on fortifying scientific inquiry so that we can understand the, sometimes remarkable, local variation in manifestation and course of mental disorders<sup>63</sup>.

## TRANSLATIONAL AND HEALTH-POLICY RESEARCH

Ensuring that populations receive high-quality care that improves mental health is the purview of policymakers. Shaping sound public policies that are based on up-to-date research can be challenging, but promising examples exist. An experimental housing policy called Moving to Opportunity found that moving from a high-poverty to a lower-poverty neighbourhood improved adult physical and mental health and subjective wellbeing over 10–15 years, despite no change in average economic status<sup>64</sup>. Moving to Opportunity was able to capitalize on the fact that public policy decisions are interconnected — it is not just health policies that influence mental health, substance use and other public health outcomes, but also economic, housing and criminal justice policies, among others<sup>65</sup>. Rapid growth in mental health and substance-use research over the past decade, as well as appeals from researchers and advocates to apply the findings in policy and practice have not yet bridged the divide between what is known and what is

done<sup>66,67</sup>. The intricacies of ensuring evidence-based health policy are not entirely understood<sup>68</sup>, but a few effective practices are being used. Advocacy organizations such as the National Alliance on Mental Illness have become trusted sources of digestible research findings<sup>69</sup>. Carefully planned links between researchers and decision makers – an approach increasingly encouraged by funders of mental health and substance-use research – can also be effective<sup>69</sup>. Such links often involve collaboration among researchers, government agencies, advocates and provider institutions to synchronize research activities with policies, health-care demands and community priorities, and to engage key stakeholders in the identification of pressing research questions and the use of study findings. In this way, policymakers have become partners in the research enterprise, helping researchers to understand what information is needed for developing or updating policies, making investment decisions, expanding access to care, improving care quality and monitoring system-level change over time. The long-term goal is that these partnerships will mobilize political will, inform policy development, and shed light on the essentials of shaping science-informed mental health and substance-use policies.

Inclusion of mental health as an explicit priority in the post-2015 development agenda (such as that included in the UN Open Working Group on Sustainable Development Goals, 2015; <https://sustainabledevelopment.un.org/content/documents/7891TRANSFORMING%20OUR%20WORLD.pdf>) provides an opportunity to mobilize the requisite political will and resources at several levels so that this ambitious agenda for research and capacity building can be realized. Lessons learned from the positive health impact as a result of Millennium Development Goals 4, 5 and 6 illuminate how multisector and multilevel cohesion of effort and commitments are powerful levers for advancing health in low-resource settings, and an opportunity for the broad community of stakeholders and advocates to improve care for individuals living with mental health and substance-use disorders.

## COLLABORATIVE CAPACITY BUILDING

New commitments and additional resources will be needed to rapidly cultivate the in-country research capacity needed to respond to the global disease burden of mental health and substance-use disorders<sup>70</sup>. The most culturally sensitive, scientifically and ethically sound, and locally relevant research requires investigators who best understand and live among the populations that they study. Funding initiatives such as the Fogarty International Center's Global Brain and Nervous System Disorders Across the Lifespan programme (<http://www.fic.nih.gov/Programs/Pages/Brain-Disorders.aspx>), the NIMH's Collaborative Hubs for International Research in Mental Health (<http://www.nimh.nih.gov/about/organization/gmh/globalhubs/index.shtml>) as well as Grand Challenges Canada's Global Mental Health granting programme (<http://www.grandchallenges.ca/grand-challenges/global-mental-health/>) that explicitly structure research capacity building into grant requirements, provide exemplary platforms to test and ultimately to systematize innovative strategies for training, mentorship and building a research culture and other infrastructural support for research in LMICs.

In addition, collaborative capacities to advance the mental health and substance-use research agenda must be developed. Capacity building in knowledge management is also integral to packaging accrued evidence so that it is accessible to policymakers and mental health technology specialists in LMICs. Platforms for knowledge sharing (for example, the Mental Health Innovation Network, <http://mhinnovation.net/>; and GHD Online, <http://www.ghdonline.org/>) can promote scientific discovery and help to harmonize the mental health and substance-use disorder research goals, processes and tools, and to catalyse the translational potential of research to policy and programmes<sup>71</sup>. Moreover, these platforms are needed to build and consolidate the community of advocates, consumers, investigators, clinicians and policymakers united in their commitment to mitigate the suffering associated with mental health and substance-use disorders, eliminate

their attendant stigma, diminish their social and economic burdens, and erase the social and health disparities perpetuated by poor access to high-quality mental health care.

## CONCLUSIONS

The formidable and rising health, economic and social burdens associated with mental health and substance-use disorders call for the prioritization of research that can inform a global response – through the development and enhancement of preventive and therapeutic strategies, health-system strengthening and policymaking – to alleviate suffering and stem the associated economic and social consequences of unmet needs. Indeed, the potential synergies among breakthroughs in basic neuroscience, epidemiological methods and implementation science, as well as the mobilization of resources and political will have generated optimism and catalysed a commitment to act among policymakers, advocates and the scientific community. Although the increase in mental health research initiatives over the past two decades are encouraging for the future challenges remain and patterns of progress have been inconsistent. We find, for example, that although response to the growing burden of depression in LMICs has led to an increase in the number of studies on effectiveness of treatments, delivery methods and task shifting to provide access to care for all populations, we do not see this same trajectory of efforts to address substance-use disorders. This occurs with the background of growing substance-use problems globally. Approaches to address substance-use disorders in LMICs are still limited, fragmented and not well vetted scientifically or culturally. On an optimistic note, the draft Social Development Goals to be passed by the UN General Assembly in September 2015 recognise mental health as integral to health and mental health is explicitly included within universal health coverage; in addition, the UN General Assembly will hold a special session on drugs in 2016. These developments have symbolic and substantive importance, and auger well for mental health within the Global Health agenda in the coming years.

1. Becker, A. E. & Kleinman, A. In *Psychiatry: Past, Present, and Prospect* (eds Bloch, S., Green, S. & Holmes, J.) (Oxford Univ. Press, 2014).
2. Murray, C. J. L. et al. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* **380**, 2197–2223 (2012).
3. Vos, T. et al. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* **380**, 2163–2196 (2012).
4. Whiteford, H. A. et al. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *Lancet* **382**, 1575–1586 (2013).
5. Charlson, F. J., Diminic, S., Lund, C., Degenhardt, L. & Whiteford, H. A. Mental and substance use disorders in sub-Saharan Africa: predictions of epidemiological changes and mental health workforce requirements for the next 40 years. *PLoS ONE* **9**, e110208 (2014).
6. Bloom, D. E. et al. *The Global Economic Burden of Non-communicable Diseases* (World Economic Forum and the Harvard School of Public Health, 2011).
7. Kleinman, A. Global mental health: a failure of humanity. *Lancet* **374**, 603–604 (2009).
8. Drew, N. et al. Human rights violations of people with mental and psychosocial disabilities: an unresolved global crisis. *Lancet* **378**, 1664–1675 (2011).
9. World Health Organization. Mental health and development: Targeting people with mental health conditions as a vulnerable group [http://apps.who.int/iris/bitstream/10665/89966/1/9789241506021\\_eng.pdf](http://apps.who.int/iris/bitstream/10665/89966/1/9789241506021_eng.pdf) (WHO, 2010).
10. World Health Organization. Global Action Plan for the prevention and control of NCDs 2013–2020 [http://www.who.int/nmh/events/ncd\\_action\\_plan/en/](http://www.who.int/nmh/events/ncd_action_plan/en/) (WHO, 2013).
11. Fricchione, G. L. et al. Capacity building in global mental health: professional training. *Harv. Rev. Psych.* **20**, 47–57 (2012).
12. Kerry, V. B. et al. US medical specialty global health training and the global burden of disease. *J. Glob. Health* **3**, 020406 (2013).
13. Tsai, A. C. et al. Global health training in US graduate psychiatric education. *Acad. Psychiatry* **38**, 426–432 (2014).
14. World Health Organization. *ATLAS on substance use. Resources for the prevention and treatment of substance use disorders* (WHO, 2010).
15. Volkow, N. D. The reality of comorbidity: depression and drug abuse. *Biol. Psychiatry* **56**, 714–717 (2004).
16. Kessler, R. C. The epidemiology of dual diagnosis. *Biol. Psychiatry* **56**, 730–737 (2004).
17. Wright, S., Gournay, K., Glorney, E. & Thornicroft, G. Dual diagnosis in the suburbs: prevalence, need, and in-patient service use. *Soc. Psychiatry Psychiatr. Epidemiol.* **35**, 297–304 (2000).

18. Moore, T. H. et al. Cannabis use and risk of psychotic or affective mental health outcomes: a systematic review. *Lancet* **370**, 319–328 (2007).
19. van Emmerik-van Oortmerssen, K. et al. Prevalence of attention-deficit hyperactivity disorder in substance use disorder patients: a meta-analysis and meta-regression analysis. *Drug Alcohol Depend.* **122**, 11–19 (2012).
20. Khantzian, E. J. The self-medication hypothesis of substance use disorders: a reconsideration and recent applications. *Harv. Rev. Psychiatry* **4**, 231–244 (1997).
21. Quello, S. B., Brady, K. T. & Sonne, S. C. Mood disorders and substance abuse disorders: a complex comorbidity. *Sci. Practice Perspect.* **3**, 13–24 (2005).
22. al'Absi, M. *Stress and Addiction: Biological and Psychological Mechanisms* (Academic, 2007).
23. UNODC. *World Drug Report 2014* (United Nations Office on Drugs and Crime, 2014).
24. Degenhardt, L., Whiteford, H. & Hall, W. D. The Global Burden of Disease projects: What have we learned about illicit drug use and dependence and their contribution to the global burden of disease? *Drug Alcohol Rev.* **33**, 4–12 (2014).
25. Balachova, T. et al. Women's alcohol consumption and risk for alcohol-exposed pregnancies in Russia. *Addiction* **107**, 109–117 (2012).
26. Edmondson, D. et al. Posttraumatic stress disorder prevalence and risk of recurrence in acute coronary syndrome in patients: a metanalytic review. *PLoS ONE* **7**, e38915 (2012).
27. Lee, A. A., McKibbin, C. L., Bourassa, K. A., Wykes, T. L. & Kitchen Andren, K. A. Depression, diabetic complications and disability among persons with comorbid schizophrenia and type 2 diabetes. *Psychosomatics* **55**, 343–351 (2014).
28. World Health Organization & Calouste Gulbenkian Foundation. *Integrating the Response to Mental Disorders and Other Chronic Diseases in Health Care Systems* (WHO, 2014).
29. Mimiaga, M. J. et al. The effect of psychosocial syndemic production on 4-year HIV incidence and risk behavior in a large cohort of sexually active men who have sex with men. *J. Acquir. Immune Defic. Syndr.* **68**, 329–336 (2014).
30. Vassileva, J. et al. Impaired decision-making in psychopathic heroin addicts. *Drug Alcohol Depend.* **86**, 287–289 (2007).
31. Sung, Y. H. et al. Decreased frontal N-acetylaspartate levels in adolescents concurrently using both methamphetamine and marijuana. *Behav. Brain Res.* **246**, 154–161 (2013).
32. National Institute of Mental Health. *Grand Challenges in Global Mental Health: Integrating Mental Health into Chronic Disease Care Provision in Low- and Middle-Income Countries* (R01) <http://grants.nih.gov/grants/guide/rfa-files/RFA-MH-13-040.html> (NIH, 2013).
33. Grimsrud, A., Stein, D. J., Seedat, S., Williams, D. & Myer, L. The association between hypertension and depression and anxiety disorders: results from a nationally-representative sample of South African adults. *PLoS ONE* **4**, e5552 (2009).
34. Wu, C. Y., Prosser, R. A. & Taylor, J. Y. Association of depressive symptoms and social support on blood pressure among urban African American women and girls. *J. Am. Acad. Nurse Pract.* **22**, 694–704 (2010).
35. Wang, P. S. et al. Use of mental health services for anxiety, mood, and substance disorders in 17 countries in the WHO world mental health surveys. *Lancet* **370**, 841–850 (2007).
36. Demyttenaere, K. et al. Prevalence, severity, and unmet need for treatment of mental disorders in the World Health Organization world mental health surveys. *J. Am. Med. Assoc.* **291**, 2581–2590 (2004).
37. Copeland, J., Thornicroft, G., Bird, V., Bowls, J. & Slade, M. Global priorities of civil society for mental health services: findings from a 53 country survey. *World Psychiatry* **13**, 198–200 (2014).
38. Saxena, S., Thornicroft, G., Knapp, M. & Whiteford, H. Resources for mental health: scarcity, inequity, and inefficiency. *Lancet* **370**, 878–889 (2007).
39. Bruckner, T. A. et al. The mental health workforce gap in low- and middle-income countries: a needs-based approach. *Bull. World Health Organ.* **89**, 184–194 (2011).
40. Araya, R., Flynn, T., Rojas, G., Fritsch, R. & Simon, G. Cost-effectiveness of a primary care treatment program for depression in low-income women in Santiago, Chile. *Am. J. Psychiatry* **163**, 1379–1387 (2006).
41. Mendenhall, E. et al. Acceptability and feasibility of using non-specialist health workers to deliver mental health care: stakeholder perceptions from the PRIME district sites in Ethiopia, India, Nepal, South Africa, and Uganda. *Soc. Sci. Med.* **118**, 33–42 (2014).
42. Patel, V. et al. Lay health worker led intervention for depressive and anxiety disorders in India: impact on clinical and disability outcomes over 12 months. *Br. J. Psychiatry* **199**, 459–466 (2011).
43. Rahman, A., Malik, A., Sikander, S., Roberts, C. & Creed, F. Cognitive behavior therapy-based intervention by community health workers for mothers with depression and their infants in rural Pakistan: a cluster-randomised controlled trial. *Lancet* **372**, 902–909 (2008).
44. World Health Organization. *Closing the Gap in a Generation: Health Equity Through Action on the Social Determinants of Health* (WHO, 2008).
45. World Health Organization. *WHO mhGAP Intervention Guide for Mental, Neurological, and Substance use Disorders in Non-Specialized Health Settings* [http://www.who.int/mental\\_health/publications/mhGAP\\_intervention\\_guide/en/index.html](http://www.who.int/mental_health/publications/mhGAP_intervention_guide/en/index.html) (WHO, 2010).
46. Patel, V. & Kim, Y.-R. Contribution of low- and middle-income countries to research published in leading psychiatry journals, 2002–2004. *Br. J. Psychiatry* **190**, 77–78 (2007).
47. Razzouk, D. et al. Scarcity and inequity of mental health research resources in low- and middle-income countries: a global survey. *Health Policy* **94**, 211–220 (2010).
48. Kieling, C. et al. Child and adolescent mental health worldwide: evidence for action. *Lancet* **378**, 1515–1525 (2011).
49. Schwarz, E. & Bahn, S. The utility of biomarker discovery approaches for the detection of disease mechanisms in psychiatric disorders. *Br. J. Pharmacol.* **153**, S133–S136 (2008).
50. Huang, J. T. et al. Independent protein-profiling studies show a decrease in apolipoprotein A1 levels in schizophrenia CSF, brain and peripheral tissues. *Mol. Psychiatry* **13**, 1118–1128 (2008).
51. al'Absi, M., Hatsukami, D. & Davis, G. Attenuated adrenocorticotrophic responses to stress predict early relapse. *Psychopharmacology* **181**, 107–117 (2005).
52. al'Absi, M., Lemieux, A. & Nakajima, M. Peptide YY and ghrelin predict craving and risk for relapse in abstinent smokers. *Psychoneuroendocrinology* **49**, 253–259 (2014).
53. Karpyak, V. M. et al. Genetic markers associated with abstinence length in alcohol-dependent subjects treated with acamprosate. *Transl. Psychiatry* **4**, e462 (2014).
54. Peedicayil, J. Epigenetic biomarkers in psychiatric disorders. *Br. J. Pharmacol.* **155**, 795–796 (2008).
55. Collins, P. Y., Patel, V. & Joestl, S. S. Grand challenges in global mental health. *Nature* **475**, 27–30 (2011).
56. van Ginneken, N. et al. Non-specialist health worker interventions for mental health care in low- and middle-income countries. *Cochrane Database Syst. Rev.* **11**, CD009149 (2013).
57. Kohrt, B. A. et al. Cultural concepts of distress and psychiatric disorders: literature review and research recommendations for global mental health epidemiology. *Int. J. Epidemiol.* **43**, 365–406 (2014).
58. Jacob, K. S. & Patel, V. Classification of mental disorders: a global mental health perspective. *Lancet* **383**, 1433–1435 (2014).
59. Chisholm, D. et al. Scale up services for mental disorders: a call for action. *Lancet* **370**, 1241–1252 (2007).
60. Eaton, J. et al. Scale up of services for mental health in low-income and middle-income countries. *Lancet* **378**, 1592–1603 (2011).
61. Lund, C. et al. Mental health is integral to public health: a call to scale up evidence-based services and developmental health research. *S. Afr. Med. J.* **98**, 444 (2008).
62. Patel, V. Why mental health matters to global health. *Transcult. Psychiatry* **51**, 777–789 (2014).
63. de Jong, J. T. Challenges of creating synergy between global mental health and cultural psychiatry. *Transcult. Psychiatry* **51**, 806–828 (2014).
64. Ludwig, J. et al. Neighborhood effects on long-term well-being of low-income adults. *Science* **337**, 1505–1510 (2012).
65. Overtveit, J. In: *Managing for Health* (ed. Hunter, D. J.) 129–148 (Routledge, 2007).
66. Mackenzie, J. *Global Mental Health from a Policy Perspective: A Context Analysis* (Overseas Development Institute, 2014).
67. Phillips, M. R. Can China's new mental health law substantially reduce the burden of illness attributable to mental disorders? *Lancet* **381**, 1964–1966 (2013).
68. Goldner, E. M., Jenkins, E. K. & Fischer, B. A narrative review of recent developments in knowledge translation and implications for mental health care providers. *Can. J. Psychiatry* **59**, 160–169 (2014).
69. Dobbins, M. et al. A randomized controlled trial evaluating the impact of knowledge translation and exchange activities. *Implement Sci.* **4**, 61 (2009).
70. Thornicroft, G., Cooper, S., Van Bortel, T., Kakuma, R. & Lund, C. Capacity building in global mental health research. *Harv. Rev. Psychiatry* **20**, 13–24 (2012).
71. Mello, M. M. et al. Preparing for responsible sharing of clinical trial data. *N. Engl. J. Med.* **369**, 1651–1658 (2014).

#### SUPPLEMENTARY MATERIAL

Is linked to the online version of this paper at: <http://dx.doi.org/10.1038/nature16032>

#### ACKNOWLEDGEMENTS

The authors thank J. Dewit, A. Garton, Y. Bodenstein and J. Nguyen at the National Institute of Mental Health for construction of the interactive map. M. A. was supported in part by the following grants: R01DA016351 and R01DA027232, and a BRAIN R21 grant (R21DA024626). F. B. was supported in part by Grand Challenges Canada Grant GMH 0094-04. We are grateful to B. Good for his insightful review and suggestions.

#### COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests. Financial support for publication has been provided by the Fogarty International Center.

#### ADDITIONAL INFORMATION



This work is licensed under the Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0>



## REVIEW OPEN

# Global research priorities for infections that affect the nervous system

Chandy C. John<sup>1</sup>, H      Carabin<sup>2</sup>, Silvia M. Montano<sup>3</sup>, Paul Bangirana<sup>4</sup>, Joseph R. Zunt<sup>5</sup> & Phillip K. Peterson<sup>6</sup>

Infections that cause significant nervous system morbidity globally include viral (for example, HIV, rabies, Japanese encephalitis virus, herpes simplex virus, varicella zoster virus, cytomegalovirus, dengue virus and chikungunya virus), bacterial (for example, tuberculosis, syphilis, bacterial meningitis and sepsis), fungal (for example, cryptococcal meningitis) and parasitic (for example, malaria, neurocysticercosis, neuroschistosomiasis and soil-transmitted helminths) infections. The neurological, cognitive, behavioural or mental health problems caused by the infections probably affect millions of children and adults in low- and middle-income countries. However, precise estimates of morbidity are lacking for most infections, and there is limited information on the pathogenesis of nervous system injury in these infections. Key research priorities for infection-related nervous system morbidity include accurate estimates of disease burden; point-of-care assays for infection diagnosis; improved tools for the assessment of neurological, cognitive and mental health impairment; vaccines and other interventions for preventing infections; improved understanding of the pathogenesis of nervous system disease in these infections; more effective methods to treat and prevent nervous system sequelae; operations research to implement known effective interventions; and improved methods of rehabilitation. Research in these areas, accompanied by efforts to implement promising technologies and therapies, could substantially decrease the morbidity and mortality of infections affecting the nervous system in low- and middle-income countries.

*Nature* 527, S178–S186 (19 November 2015), DOI: 10.1038/nature16033

This article has not been written or reviewed by *Nature* editors. *Nature* accepts no responsibility for the accuracy of the information provided.

**R**ecent improvements in the detection of infectious organisms that can affect the nervous system has led to the realization that a substantial proportion of chronic neurological, cognitive and behavioural disease may actually have an acute and preventable origin. Infectious organisms may infect the nervous system directly, as in rabies and bacterial meningitis, or may cause neurocognitive disorders in the absence of direct infection of the nervous system, as in malaria or hookworm infection. This Review identifies global research priorities for infections that affect the nervous system, with the ultimate goal of stimulating research in these priority areas to substantially reduce morbidity associated with nervous system infections worldwide.

## METHODS

For this Review, we chose illustrative infections that cause considerable nervous system morbidity in children and adults in low- and middle-income countries (LMICs). These infections are examples and are not meant to be exhaustive. Estimates of infection global frequency and types of nervous system involvement were obtained through PubMed searches using the infection name and were accompanied by any of the following terms: neurologic, nervous system, cognition, cognitive, development, neurodevelopment, impairment, deficit, sequelae, brain injury, brain damage, mental health, behavioral or neuropathy. If available, World Health Organization (WHO) documents were also reviewed for each disease. The authors came to a consensus on the key

research priority areas, on the basis of a literature review and research experience.

## INFECTIONS AFFECTING THE NERVOUS SYSTEM

The global distribution, frequency and types of neurological, cognitive and mental health disorders associated with key infections are presented in Table 1. For classification purposes, infections are reviewed according to type of microorganism (virus, bacteria, fungus or parasite) in the sections that follow. However, microorganisms within a group (for example, the viruses HIV and rabies) can affect the nervous system in as varied a manner as microorganisms of different groups (for example, the virus HIV and the malaria-causing parasite *Plasmodium falciparum*).

### Viral infections

Worldwide, rabies and Japanese encephalitis virus (JEV) are responsible for an estimated annual mortality of 60,000 and 17,000 people, respectively<sup>1,2</sup>. Cases of JEV encephalitis are restricted to Asia, whereas rabies is a scourge throughout Southeast Asia, Africa and Latin America and occurs, although less frequently, in other areas worldwide. Rabies and herpes simplex virus (HSV) encephalitis, which is also present worldwide, lead to high mortalities without treatment<sup>3</sup>. JEV has variable mortality, depending, in part, on the infected individual's age. Among survivors, long-term cognitive or neurological impairment is

<sup>1</sup>Ryan White Center for Pediatric Infectious Disease and Global Health, Department of Pediatrics, Indiana University School of Medicine, Indianapolis, Indiana 46202, USA. <sup>2</sup>Department of Biostatistics and Epidemiology, College of Public Health, University of Oklahoma Health Sciences Center, Oklahoma City, Oklahoma 73104, USA. <sup>3</sup>Department of Bacteriology, US Naval Medical Research Unit No. 6, Lima, Peru. <sup>4</sup>Department of Psychiatry, Makerere University College of Health Sciences, Kampala, Uganda. <sup>5</sup>Department of Epidemiology, University of Washington, Seattle, Washington 98195, USA. <sup>6</sup>Division of Infectious Diseases and International Medicine, University of Minnesota, Minneapolis, Minnesota 55455, USA. Correspondence should be addressed to C. C. J. e-mail: chjohn@iu.edu.

present in as many as 70% of those with HSV encephalitis<sup>4</sup> and 30–50% of those with JEV encephalitis<sup>5</sup>. Most cases of JEV infection (as opposed to encephalitis) are asymptomatic or mildly symptomatic and require no treatment<sup>6</sup>, but the long-term neurocognitive consequences of asymptomatic or mildly symptomatic JEV infections are unknown.

Varicella zoster virus (VZV), the cause of chickenpox, can, like other herpesviruses, establish a latent infection. Reactivation is usually owing to suppression of cell-mediated immunity, most commonly age-related immunosenescence. Central nervous system (CNS) reactivation is relatively uncommon, but reactivation in a dorsal root ganglion can lead to herpes zoster, which is associated with debilitating chronic pain<sup>7</sup>. Herpes zoster is the single most common infection of the nervous system, with an estimated one million new cases each year in the United States alone<sup>8</sup>. There is little data on nervous system VZV infection in LMICs.

Congenital cytomegalovirus (CMV) infection is the most common acquired cause of hearing loss in children in the United States<sup>9</sup>. The percentage of the population testing positive for CMV is higher in LMICs than in high-income countries, but the incidence of congenital CMV infection and of symptomatic CNS disease and hearing loss in most LMICs remains almost unknown<sup>10</sup>.

Dengue, and to a lesser extent chikungunya, viruses will probably become leading global causes of arboviral encephalitis in the next decade<sup>2,11,12</sup>. Although encephalitis is present in only a fraction of those infected with dengue virus, the large number of dengue infections worldwide could lead to hundreds of thousands of encephalitis cases (Table 1).

## HIV and opportunistic infections

In 2012, an estimated 2.3 million people were infected with HIV, and 1.6 million died of AIDS-related illnesses worldwide<sup>13</sup>. HIV-associated neurological syndromes are classified as primary HIV infection, secondary or opportunistic infection, and treatment-related neurological disease. Most often, primary HIV infection causes acute aseptic (viral) meningitis or meningoencephalitis (MEC). HIV-associated neurocognitive disorder (HAND) is a neurodegenerative condition characterized by cognitive, motor and behavioural abnormalities that is becoming more common with an increase of HIV in people older than 50 years<sup>14</sup>. In LMICs, where only one third of patients requiring highly active antiretroviral therapy (HAART) receive it, the opportunistic infections cryptococcal meningitis, tuberculous meningitis, cerebral toxoplasmosis, progressive multifocal leukoencephalopathy and CNS cytomegalovirus infection remain common<sup>15</sup>.

The widespread implementation of combination antiretroviral therapy (cART) has changed the presentation, manifestation and epidemiology of many conditions and opportunistic infections, owing to immune reconstitution syndrome and increased prevalence of cognitive impairment and neuropathy with additional morbidities in patients who are now living longer<sup>16</sup>. The epidemiology and neurological outcomes of HIV infection are also affected by underlying malnutrition and variations in endemic pathogens.

Cryptococcal meningitis is a leading cause of mortality in LMICs where access to antiretroviral therapy is limited<sup>17</sup>. Most cases occur in sub-Saharan Africa, followed by South and Southeast Asia<sup>18</sup>. Seroprevalence for *Toxoplasma gondii* in people with HIV ranges from 10% to 80% with the highest proportions in African countries<sup>19</sup>, and cerebral toxoplasmosis is the most common cerebral mass lesion in patients with AIDS.

## Bacterial infections

The most common bacterial infections affecting the nervous system are sepsis and meningitis in neonates; bacterial meningitis due to *Streptococcus pneumoniae*, *Haemophilus influenzae* type b and *Neisseria meningitidis* in children and adults; and tuberculous meningitis in children and adults.

Neonatal meningitis and neonatal sepsis are associated with long-term neurological and cognitive impairment<sup>20</sup>; primarily impairment of hearing, vision or motor function; cerebral palsy; and epilepsy. In

LMICs, it is estimated that 23% of neonates who have survived meningitis sustain moderate to severe neurodevelopmental impairment<sup>21</sup>. *Staphylococcus aureus*, Gram-negative infections, including *Escherichia coli*, *Klebsiella pneumoniae*, *Acinetobacter*, non-typhoidal *Salmonella* and group B *Streptococcus*, are the leading causes of neonatal sepsis and meningitis in most LMICs<sup>22–24</sup>. Recent reports from Africa and India suggest an alarming increase in drug resistance among Gram-negative organisms infecting neonates<sup>24,25</sup>.

In children and adults in high-income countries, bacterial meningitis due to *S. pneumoniae*, *H. influenzae* type b and *N. meningitidis* has decreased dramatically following immunization with conjugate vaccines<sup>26</sup>. However, availability of these vaccines in LMICs is variable, and bacterial meningitis still affects 1.2 million individuals annually<sup>26</sup>, causing neurocognitive sequelae in 23% of affected children. Steroids as an adjunctive therapy have reduced neurological sequelae in high-income countries, particularly in adults, but have shown no benefit in LMICs<sup>27</sup>. Factors such as organism strain, co-infections, adjunctive and supportive therapies, and underlying conditions such as poor nutrition that can affect immune response often differ between high-income countries and LMICs, and may play a part in the different results seen in some clinical trials.

Tuberculous meningitis (TBM) occurs in around 1% of all cases of tuberculosis, but is the most severe form of extrapulmonary tuberculosis, resulting in death or severe disability in about 50% of those with the disease<sup>28</sup>. Bacterial meningitis, and particularly TBM, can result in hydrocephalus, which is often difficult to treat in LMICs because neurosurgeons are typically not available and the supplies for ventriculoperitoneal shunt placement are often not present or very limited.

The WHO estimated that approximately 10.6 million new cases of syphilis occurred in 2008 (ref. 29), but precise estimates of the incidence of neurosyphilis are not available.

## Parasitic infections

Malaria in humans is caused by one of five *Plasmodium* species, but neurological disabilities are most frequently associated with *Plasmodium falciparum*. Although *P. falciparum* does not directly infect brain tissue, severe infection can lead to coma. One in four children with cerebral malaria develops long-term cognitive impairment<sup>30</sup>, and recent studies suggest that children with severe malarial anaemia also have long-term cognitive impairment<sup>31</sup>. Behavioural problems and epilepsy are other long-term consequences of cerebral malaria<sup>32</sup>. Children with repeated episodes of uncomplicated malaria have motor and cognitive problems<sup>33</sup>. The mechanisms by which malaria leads to neurocognitive problems are not fully defined, and the neurocognitive burden of malaria owing to other *Plasmodium* species has not been characterized.

Neurocysticercosis, which is endemic in areas with poor pig management practices and sanitation, occurs when the larval stages of *Taenia solium* infect the brain. In LMICs in which neurocysticercosis is endemic, it is the leading identified cause of seizures. The proportion of larval infections migrating to the brain is unknown, but some individuals with neurocysticercosis never show neurological symptoms<sup>34</sup>. Neurocysticercosis can be fatal, most often following complications of surgery to treat the hydrocephalus associated with intraventricular or subarachnoid neurocysticercosis, but overall mortality estimates have been difficult to obtain<sup>35</sup>.

Neuroschistosomiasis results from the migration of schistosome eggs or worms to the brain or spinal cord, and may occur following infection with *Schistosoma japonicum*, *Schistosoma mansoni* or *Schistosoma haematobium*. Brain involvement may occur in the acute phase (acute schistosomal encephalopathy) or in the chronic phase (cerebral schistosomiasis or pseudotumoral encephalic schistosomiasis). The spinal cord may also be involved, often leading to hemiparesis<sup>36</sup>. CNS symptoms and epilepsy are reported to occur in 2.6% and 2.1% of *S. japonicum* infections, respectively<sup>37</sup>. Neuroschistosomiasis can be fatal, especially in its tumour-like form when it affects the cerebellum, but accurate mortality rates are unavailable<sup>38</sup>.

**Table 1** | Neurocognitive and mental health consequences of major infectious diseases that affect the nervous system.

Infectious disease	Regions affected	Estimated prevalence or annual incidence of infection	Health consequences		
			Neurological	Cognitive	Mental health
VIRAL					
Arboviruses					
Dengue virus	Global, most common in South Asia, Africa and Latin America	390 million (95% CI, 284–528 million)	<ul style="list-style-type: none"><li>• Meningitis, meningoencephalitis, encephalitis, seizures, Guillain–Barré syndrome, neuralgic amyotrophy, hypokalaemic paralysis, and dengue myositis</li><li>• In one cohort, dengue had neurological manifestations in 9.3% of children and adults</li><li>• There is limited information about long-term sequelae in dengue, but there is evidence of significant long-term neurological complications</li></ul>	<ul style="list-style-type: none"><li>• Not studied</li></ul>	Case reports of mania and depression
Chikungunya virus	Global, most common in South Asia, Africa and Latin America	33,000–93,000	<ul style="list-style-type: none"><li>• Encephalitis, febrile seizures, meningismus, myelopathy or myeloneuropathy</li></ul>	<ul style="list-style-type: none"><li>• Not studied</li></ul>	Not studied
Japanese encephalitis	Southeast Asia	35,000–50,000	<ul style="list-style-type: none"><li>• CNS complications during the acute illness include delirium, seizures, axial rigidity, extrapyramidal signs, cranial nerve palsies, ataxia, paraplegia and segmental sensory disturbances</li></ul>	<ul style="list-style-type: none"><li>• Among survivors, 30–50% have significant neurological, cognitive or psychiatric sequelae</li></ul>	Among survivors, 30–50% have significant neurological, cognitive, or psychiatric sequelae
Rhabdoviruses					
Rabies	Global, greatest in sub-Saharan Africa, Southeast Asia and Latin America	60,000 (probably an underestimate)	<ul style="list-style-type: none"><li>• Severe encephalitis, which is almost 100% fatal</li></ul>	<ul style="list-style-type: none"><li>• Fatal</li></ul>	Fatal
Herpesviruses					
HSV encephalitis	Global	Present in all countries where HSV testing has been performed, but no reliable global estimates	<ul style="list-style-type: none"><li>• If untreated, as in most LMICs, there is a high fatality rate for HSV-1 (around 70%), lower (around 15%) if treated. Long-term neurological complications occur in around 70% of adult survivors, including seizure disorder and hemiparesis. In one cohort, neurological sequelae occurred in 63% of infections in children, including seizures in 44% and developmental delays in 25%</li></ul>	<ul style="list-style-type: none"><li>• In one study of adult survivors, long-term cognitive sequelae included memory impairment (69%)</li></ul>	Personality or behavioural impairment in 45% of adult survivors
VZV	Global	No reliable global estimates	<ul style="list-style-type: none"><li>• CNS: stroke, meningoencephalitis, myelitis</li><li>• PNS (more common): herpes zoster with chronic pain</li></ul>	<ul style="list-style-type: none"><li>• Very limited studies with conflicting results</li></ul>	Major depression
Congenital cytomegalovirus	Global	0.6–0.7% of live births in high-income countries and 1–5% of live births in LMICs	<ul style="list-style-type: none"><li>• Most common non-hereditary cause of hearing loss in children in the United States</li><li>• There are no reliable estimates for frequency of hearing loss due to the infection in most LMICs</li></ul>	<ul style="list-style-type: none"><li>• Symptomatic infection, seen in 10–15% of congenitally infected children, is associated with significant global developmental delay in around 50% of affected children</li></ul>	Behavioural problems
HIV-related					
HIV	Global, greatest burden in sub-Saharan Africa and Asia	Annual incidence estimate is 2.3 million (95% CI, 1.9–2.7 million) with 34 million people living with HIV/AIDS worldwide, of whom 23 million live in sub-Saharan Africa and 3.5 million live in Southeast Asia	<ul style="list-style-type: none"><li>• HIV associated opportunistic infections, aseptic meningitis, AIDS encephalopathy, Bell’s palsy, progressive multifocal leukoencephalopathy, primary CNS lymphoma, stroke, transverse myelitis, HIV-associated peripheral neuropathy, inflammatory demyelinating polyneuropathy, immune reconstitution inflammatory syndrome and vacuolar myelopathy</li></ul>	<ul style="list-style-type: none"><li>• Asymptomatic neurocognitive impairment, mild neurocognitive disorder and HIV-associated dementia</li></ul>	Delirium, major depression, bipolar disorder (including AIDS mania), schizophrenia, substance abuse or dependence and post-traumatic stress disorder
Cryptococcal meningitis	Global, greatest burden in sub-Saharan Africa and Asia	Annual incidence estimate: 957,900 in 2009, approximately 624,700 deaths annually	<ul style="list-style-type: none"><li>• Headache, meningismus, intracranial hypertension, mental status changes, focal intracerebral granulomas (cryptococcomas), hydrocephalus (communicating and non-communicating), papilledema, sensorineural deafness, cranial nerve palsies, motor and sensory deficits, cerebellar dysfunction and seizures</li></ul>	<ul style="list-style-type: none"><li>• Mimicking of vascular dementia, and reversible dementia</li></ul>	Personality change, confusional psychosis and mania
Toxoplasma encephalitis	Global, greatest burden in sub-Saharan Africa and Asia	No reliable global estimates of incidence of toxoplasma encephalitis, but toxoplasma infection is present in 14% of the population in the United States, compared with 23–47% in some European, Latin American and African countries	<ul style="list-style-type: none"><li>• Headache, focal neurological deficit, seizures and altered mental status</li></ul>	<ul style="list-style-type: none"><li>• Dementia</li></ul>	Schizophrenia and behaviour disorders



BACTERIAL					
Neonatal sepsis and meningitis	Global	Annual incidence estimates for south Asia, sub-Saharan Africa and Latin America: neonatal sepsis, 1.7 million (uncertainty estimate, 1.1–2.4 million); neonatal meningitis, 200,000; 95% CI, 21,000–350,000	<ul style="list-style-type: none"> <li>Little data for neonatal sepsis globally, especially among those more than 32 weeks gestation or more than 1,500g</li> <li>23% (95% CI, 19–26%) of neonatal meningitis survivors (or 18,000 children; 95% CI, 2,700–35,000) estimated to sustain moderate to severe neurodevelopmental impairment</li> <li>In sepsis or meningitis, the primary neurological sequelae are cerebral palsy, impairment to vision, hearing and motor function, and seizure disorders</li> </ul>	Limited studies reporting cognitive impairment; developmental delay or learning difficulties are frequent in sepsis (30.0%; IQR, 26.4–44.4%) and meningitis (33.3%; IQR, 26.7–36.8%)	No data
Bacterial meningitis	Global	Annual incidence estimate: 1.2 million	<ul style="list-style-type: none"> <li>22.8% (IQR, 12.1–29.2%) have at least 1 neurocognitive sequela at discharge, 19.9% (IQR, 12.1–35.2%) have at least 1 sequela post-discharge; 16.0% (IQR, 7.1–21.2%) have at least 1 major sequela at discharge, 12.8% (IQR, 7.1–21.1%) have at least 1 major sequela post discharge</li> <li>Neurological sequelae include motor deficits, hearing loss and visual disturbances</li> <li>Risk of major sequelae is higher in Africa (25.1%) and southeast Asia (21.6%) compared with Europe (9.4%)</li> </ul>	In children, cognitive impairment including low IQ, academic limitations, and impaired executive function and in adults, cognitive impairment with slower cognitive speed seen	Behavioural changes and emotional disturbance including ADHD and learning difficulties
Tuberculous meningitis (also an opportunistic infection in HIV)	Global, most burden in sub-Saharan Africa and Asia	No reliable global incidence estimates; highest in countries with high prevalence of HIV infection	<ul style="list-style-type: none"> <li>Neurological sequelae in 53.9% of child survivors (95% CI, 42.6–64.9)</li> <li>Gross and fine motor impairment in children</li> <li>Motor deficits, optic atrophy, ophthalmoplegia, and hearing impairment in adults and older children</li> </ul>	Cognitive impairment in all areas tested, and poor scholastic progress	Emotional disturbance
Neurosyphilis	Global	No reliable global incidence estimates; most cases occur in HIV-positive individuals	<ul style="list-style-type: none"> <li>Meningitis, cerebrovascular infarction, and paresis, tabes dorsalis (ataxia, paraesthesia and bladder dysfunction)</li> </ul>	Impaired memory, disorientation and dementia	Dementia, depression, delirium, mania and psychosis
PARASITIC					
Neurocysticercosis	Global, greatest burden in pig-raising areas with poor sanitation	2010 prevalence estimate: 1.4 million (95% CI, 1.3–1.6 million) (epilepsy only)	<ul style="list-style-type: none"> <li>Among people with symptomatic neurocysticercosis diagnosed with brain imaging: seizures and epilepsy (78.8%; 95% CI, 65.1–89.7%), headaches (37.9%; 95% CI, 23.3–53.7%), focal deficits (16.0%; 95% CI, 9.7–23.6%) and symptoms associated with increased intracranial pressure (11.7%; 95% CI, 6.0–18.9%)</li> </ul>	<ul style="list-style-type: none"> <li>Case reports of cognitive decline</li> <li>Cognitive symptoms of neurocysticercosis with active cysts: affects naming, verbal fluency and non-verbal memory</li> </ul>	Neurocysticercosis with active cysts: dementia (12.5%) and cognitive impairment, but not dementia (27.5%); psychosis
Malaria	Sub-Saharan Africa, Latin America, Asia and Oceania	Annual incidence estimate: 216 million	<ul style="list-style-type: none"> <li>Cerebral malaria: 5–28% of children have neurological deficits on discharge, including epilepsy, acute hemiparesis, hypertonia, cortical blindness and ataxia</li> <li>By 6-month follow-up the percentage of children with deficits has decreased to 0–4.4%, primarily in the areas of gross motor and fine motor skills</li> <li>Uncomplicated malaria: motor skills</li> </ul>	<ul style="list-style-type: none"> <li>Cerebral malaria affects general cognition, attention, working memory, visual spatial skills, somatosensory discrimination, speech and language, and receptive and expressive language</li> <li>Thirteen IQ point difference from non-affected children 1 year after episode, and around 26% of children have impairment 2 years after</li> <li>Severe malaria with neurological involvement affects executive function</li> <li>Severe malarial anaemia affects overall cognition estimated to lead to the equivalent of an 11 IQ point difference from non-affected community children</li> <li>Malaria with multiple seizures leads to speech and language problems</li> <li>Malaria with impaired consciousness leads to attention/language problems</li> <li>Uncomplicated malaria leads to language problems</li> <li>Asymptomatic malaria leads to problems with fine motor coordination, attention and abstract reasoning</li> </ul>	Cerebral malaria: internalizing and externalizing problems, ADHD, disruptive behaviour, psychosis and depression
STH infection	Global, greatest burden in sub-Saharan Africa and Southeast Asia	Estimated 2010 prevalence: hookworm infected 439 million (95% CI, 406–480), <i>Ascaris lumbricoides</i> infected 819 million (95% CI, 772–892) and <i>Trichuris trichiura</i> infected 465 million (95% CI, 430–508).	<ul style="list-style-type: none"> <li>Not described</li> </ul>	<ul style="list-style-type: none"> <li>School-aged children: <i>T. trichiura</i> and <i>A. lumbricoides</i> affected learning and verbal memory in one study</li> <li>In another, general STH infection reduced memory capacity, rate of processing and attention</li> </ul>	Children under 5 years of age: social and emotional disturbances (combined with anaemia)
Schistosomiasis	Global, greatest in sub-Saharan Africa and Southeast Asia	Estimated 2010 prevalence: 252 million infected	<ul style="list-style-type: none"> <li>Acute schistosomal encephalopathy: headache, confusion, seizure, loss of consciousness, focal deficits, visual impairment and ataxia</li> <li>Cerebral schistosomiasis: headaches, motor deficits, visual abnormalities, seizures, altered mental status, vertigo, sensory impairment, speech disturbances and ataxia</li> <li>Spinal cord schistosomiasis: lower limb weakness, bladder dysfunction, lower limb paraesthesia, hypoaesthesia/anaesthesia, deep tendon reflex abnormalities, constipation and impotence in 80% of cases</li> </ul>	For <i>Schistosoma japonicum</i> infection in children (not neurological infection): verbal memory and verbal fluency affected	No data

ADHD, attention deficit disorder; CI, confidence interval; CNS, central nervous system; HSV, herpes simplex virus; IQR, interquartile range; LMICs, low- and middle-income countries; PNS, peripheral nervous system; STH, soil-transmitted helminths; VZV, varicella-zoster virus. Prevalence estimates are typically used (for example, STH infections and schistosomiasis) because accurate incidence numbers for these infections are difficult to obtain.

**Table 2** | Potential areas for intervention in infectious diseases that affect the nervous system.

Disease	Vaccine available	Control of zoonotic reservoirs	Control of vector populations	Treatment
<b>VIRAL</b>				
Dengue	New dengue vaccines being tested in large field trials	NA	Yes	None available
Chikungunya	No	NA	Yes	None available
Japanese encephalitis	Yes	No	Yes	None available
Rabies	Yes	Yes	NA	None available
HSV encephalitis	No	NA	NA	Yes
VZV	Yes	NA	NA	Yes
Congenital cytomegalovirus	No	NA	NA	Yes
<b>HIV-related</b>				
HIV	No	NA	NA	Yes
Cryptococcal meningitis	No	NA	NA	Yes
Toxoplasma encephalitis	No	Yes	NA	Yes
<b>BACTERIAL</b>				
Neonatal sepsis and meningitis	No	NA	NA	Yes
Bacterial meningitis	Yes, for <i>Haemophilus influenzae</i> type b, and pneumococcal (multiple serotypes) and meningococcal (A, C, Y and W135) meningitis	NA	NA	Yes
Tuberculous meningitis	Partial protection provided by BCG vaccination	Infrequent (cases due to <i>Mycobacterium bovis</i> and <i>Mycobacterium caprae</i> , both of which are present in cattle, reported)	NA	Yes
Neurosyphilis	No	NA	NA	Yes
<b>PARASITIC</b>				
Neurocysticercosis	No	Porcine vaccine trials underway, pig treatment available	NA	Yes
Malaria	RTS,S vaccine had efficacy in phase III studies and other vaccines are being developed	NA except for <i>Plasmodium knowlesi</i>	Yes	Yes
STH	No. Hookworm vaccine is in phase I trials, but is linked to adverse events	NA except for <i>Toxocara canis</i>	NA	Yes
Schistosomiasis	No, but phase I vaccine trials are ongoing	Bovine vaccine trials underway for <i>Schistosoma japonicum</i>	Yes	Yes

BCG, Bacillus Calmette–Guérin; HSV, herpes simplex virus; NA, not applicable; STH, soil-transmitted helminth; VZV, varicella-zoster virus.

Soil-transmitted helminths (STH) affect millions, most frequently children. Determining the part played by STH in cognitive impairment of children is complicated because STH infection is associated with many confounders, but data that support a role for STH in neurobehavioural outcomes include a study showing that infants and children under 5 years of age with anaemia and STH infection show disturbed social and emotional behaviour. Another study showed that treating school-aged children with antiparasitic drugs and iron supplementation improved attention, memory and processing speed<sup>39</sup>.

## GLOBAL RESEARCH PRIORITIES

Prevention of infections that affect the nervous system is the highest research priority, as complete prevention of infection removes all risk of nervous system sequelae. However, treatment of nervous system sequelae and rehabilitation of individuals with nervous system morbidity are also important for the millions who currently live with the nervous system effects of infections. Prevention and treatment of infections that affect the nervous system requires the identification of the pathogens responsible, the pathogen reservoirs and the potential points at which the pathogen life cycle can be interrupted. Table 1 lists specific pathogens, their known nervous system manifestations, and current knowledge gaps regarding incidence and long-term sequelae for each pathogen. Table 2 outlines whether specific interventions (vaccines, control of zoonotic reservoirs or vector populations, and treatment) are available for each pathogen. Table 3 provides a summary of global research priorities for infections that affect the nervous system. These research priorities are discussed in more detail below.

## Diagnosis

Improved diagnosis lies at the heart of all research priorities for infections that affect the nervous system because all other research areas

depend on accurate infection diagnosis. Improved diagnosis requires better tests to detect infection, better clinical diagnostic algorithms to detect infection and better tools to assess the nervous system effects of infection, including cognitive and mental health sequelae.

Affordable, easy-to-use, rapid diagnostic assays — preferably point-of-care — that can identify infections affecting the nervous system are a high priority. This includes diagnostic tests for infections that directly infect nerve cells and those that do not (for example, malaria and STH). For infections that affect the CNS, field diagnoses are needed to identify when the infection has entered the CNS. Serological assays are available to detect schistosomiasis<sup>36</sup> and cysticercosis, but these tests are not specific for CNS infection, and the blood-brain barrier may prevent the detection of antigens in serum<sup>40</sup>. In the case of bacterial, fungal or viral CNS infections, although lumbar puncture to obtain cerebrospinal fluid is a routine procedure at many centres in LMICs, most lack the capacity for standard bacterial, fungal or viral cultures, let alone more sophisticated testing such as PCR, which is essential for detecting many viral infections. Even in high-income countries with advanced molecular diagnostics, an aetiological agent is identified in less than half of individuals with encephalitis. Because many cases of idiopathic encephalitis are probably caused by viruses still to be characterized, the development of metagenomic and high-throughput screening techniques for viral detection is a research priority, with the goal of eventually developing low-cost diagnostic point-of-care assays for the pathogens identified. For certain CNS infections, notably neurocysticercosis, neuroschistosomiasis, CNS tuberculosis and CNS toxoplasma infection, neuroimaging with CT scans or MRI is needed to make a diagnosis. Although availability of neuroimaging is becoming more widespread, many facilities in LMICs still lack these costly imaging modalities, underscoring the need for research on simple, accurate, low-cost, point-of-care diagnostic tests for detecting infections that affect the nervous system.

One example of how improved diagnostic tools can have an impact is a study<sup>41</sup> from India in which a simple diagnostic algorithm and basic treatment for neonatal sepsis, all performed by village health workers, led to a 63% reduction in neonatal mortality among preterm infants<sup>41</sup>. Simple algorithms for other infections that affect the nervous system, coupled with the ability to provide effective therapy following diagnosis, or appropriate referral for screening algorithms, have the potential to substantially reduce morbidity and possibly mortality from these infections. Improved cross-cultural measurements of neurodevelopment and mental health are a key research priority, and reviewed in the article in this collection on child neurodevelopment (see page S155).

## Epidemiology and primary prevention

The lack of affordable, non-invasive, rapid diagnostics for infection and nervous system effects of infection limits our ability to quantify the burden of infection-related nervous system disability (Table 1). Well-designed studies of disease epidemiology are also required for the accurate measurement of disease incidence, and of the type and duration of nervous system sequelae of infection.

A challenge to the estimation of infection-related nervous system disease is that the symptoms these infections result in, such as epilepsy, hemiparesis or cognitive impairment, are included as 'chronic diseases' in global burden estimates. Careful epidemiological assessment could lead to the more accurate attribution of a portion of 'chronic disease' to its infectious component. For example, the *Global Burden of Disease Study 2010* (ref. 42) attributed some of the disability-adjusted life years of epilepsy to neurocysticercosis. This infection is also associated with stroke<sup>43</sup>, which was ranked third in terms of disability-adjusted life years in 2010 (ref. 42), but none of this burden was attributed to neurocysticercosis owing to lack of data. Thus the true burden of nervous system disease owing to neurocysticercosis was probably significantly underestimated.

The most cost-effective method of preventing infection is immunization, discussed in the section on vaccines below. For infections for which there is no immunization, or for which immunization is not highly successful, research is required on sustainable preventive methods. For vector-borne illness, for example, insecticide-based interventions such as insecticide-treated bed nets have reduced malaria incidence and mortality in many areas<sup>44</sup>. But increasing pyrethroid resistance<sup>45</sup> highlights the need for ongoing research even for interventions with documented past success.

## Pathogenesis

Disease pathogenesis may be the most neglected research focus of infection-related nervous system disease in LMICs. Although some studies on the pathogenesis of infection-related nervous system disease in individuals in high-income countries are available<sup>46,47</sup>, far fewer studies of pathogenesis have been conducted in individuals from LMICs. Even in high-income countries, studies of infection pathogenesis often use animal models, which may incompletely recapitulate the host response in humans. The host immune response probably contributes to both defence against invading pathogens and subsequent damage to the nervous system<sup>48</sup>, but the type and role of specific cells in the immune response at different infection stages are poorly described. Similarly, it is often unclear which antigens or components of the infecting organism confer neurovirulence. The roles of innate immunity, the microbiome, and co-infection with endemic pathogens, including HIV, in contributing to infection-related nervous system disease are also poorly understood (Box 1). Without an understanding of the pathogenesis of infection-related nervous system disease, it is difficult to rationally plan for adjunctive interventions to prevent or reduce nervous system injury. Although adjunctive interventions have been elusive, those proven successful (for example, steroid treatment in tuberculous meningitis<sup>49</sup>) have made a major difference in improving neurocognitive and behavioural outcomes. An understanding of the development and types

**Table 3 | Global research for infections that affect the nervous system.**

Priority area	Research needed
<b>Diagnosis</b>	<ul style="list-style-type: none"> <li>• Rapid, accurate, low-cost, point-of-care diagnostic tests for infections that affect the nervous system</li> <li>• Clinical diagnostic algorithms for infections that affect the nervous system</li> <li>• Improved testing for detection of infection-related nervous-system disabilities</li> </ul>
<b>Epidemiology</b>	<ul style="list-style-type: none"> <li>• Accurate incidence and prevalence estimates of common infections that affect the nervous system</li> <li>• Accurate identification and frequency estimates of nervous-system manifestations and sequelae</li> <li>• Identification of potentially modifiable risk factors specific to infections that affect the nervous system</li> </ul>
<b>Pathogenesis</b>	<ul style="list-style-type: none"> <li>• Identification of host response pathways that lead to nervous-system deficits or to clinical immunity</li> <li>• Identification of pathogen factors that lead to nervous-system deficits or to clinical immunity</li> <li>• Assessment of risks and interactions of co-infections and co-morbidity</li> </ul>
<b>Vaccine development</b>	<ul style="list-style-type: none"> <li>• Develop safe and effective vaccines based on immunology, epidemiology and pathogenesis studies</li> <li>• Phase I and II trials</li> <li>• Phase III trials</li> </ul>
<b>Treatment</b>	<ul style="list-style-type: none"> <li>• Effective adjunctive treatment to prevent or decrease nervous-system deficits or disabilities</li> <li>• Low cost, low toxicity antimicrobials that work against drug-resistant pathogens</li> <li>• Multi-site, large clinical trials that provide definitive answers on interventions</li> <li>• Effective or improved primary treatment of infection</li> </ul>
<b>Rehabilitation</b>	<ul style="list-style-type: none"> <li>• Effective and feasible physical, occupational and cognitive rehabilitation programmes</li> </ul>
<b>Operations and implementation</b>	<ul style="list-style-type: none"> <li>• Optimal methods to implement or operationalize interventions with known efficacy</li> </ul>

of protective immune responses to antigens or antigenic variants of a pathogen is also fundamental to the development of vaccines, which are, in most cases, the most cost-effective method of preventing infection.

## Vaccine development

Vaccines are available to prevent the neurological complications of measles, mumps, rubella, poliomyelitis and varicella virus as well as *H. influenzae* type b, *S. pneumoniae* and *N. meningitidis*. Effective vaccines are also available for rabies and Japanese encephalitis. Together, these vaccines have saved millions of lives as well as prevented long-term nervous system complications in millions of children and adults.

Research priorities for vaccine development include the utilization of disease immunology, epidemiology and pathogenesis studies to develop safe, effective vaccines, and the performance of phase I, II and III trials to determine vaccine efficacy and safety in humans. In *P. falciparum* malaria, for example, knowledge of antibody and T-cell immune responses to circumsporozoite protein (CSP)<sup>50,51</sup> led to phase I and II trials of the CSP-based RTS,S vaccine<sup>52</sup>. These successful trials led to the recently completed phase III trials of RTS,S<sup>53</sup>. This constituted a major advance in the vaccine field because they established RTS,S as the first successful vaccine in humans against a parasite. However, the relatively modest efficacy (30–50%) of RTS,S was not surprising in light of the known complexity of the human immune response to *P. falciparum* in endemic populations. Hence, work continues on the development of more effective vaccines. Understanding the human immune response to *P. falciparum* infection will be key to the development of vaccines with improved efficacy and safety.

## Treatment

Treatment with antimicrobials is designed to clear infection or reduce infectious load, decrease disease severity, and ideally to provide a degree of secondary prevention against the nervous system effects of the infection. For viral infections, with the exception of HSV and HIV, there is often no specific treatment. Even for HSV encephalitis, standard treatment (intravenous acyclovir) is unavailable in many parts of the world. Cost effectiveness and stakeholder analyses could be useful in influencing policymakers to increase availability of antiviral



## BOX 1 | EMERGING RESEARCH AREAS

### Diarrhoeal disease

Diarrhoea is a leading cause of mortality in children living in low- and middle-income countries (LMICs)<sup>62</sup>. Most of these deaths occur in Africa, Southeast Asia and in eastern Mediterranean countries<sup>63</sup>. In 2010, it is estimated that there were 1.731 billion episodes of diarrhoea worldwide of which 36 million progressed to severe diarrhoea and 700,000 episodes resulted in death<sup>64</sup>. Its occurrence in the first 2 years of life is associated with an 8 cm decrease in height and a 10 point drop in IQ by the time children are around 7 to 9 years old<sup>65</sup>. The mechanism by which diarrhoea affects cognition is not clear, but it could be through the effect of diarrhoea on stunting, which in turn predicts future cognition<sup>66</sup>. However, during diarrhoea-free periods in the first 2 years of life, children experience catch-up growth and may return to their original growth trajectories<sup>67</sup>. This highlights the importance of effective interventions for diarrhoea to sustain the child's developmental potential. The high frequency of diarrhoea episodes during this critical developmental stage and the large number of cases makes diarrhoea a major public health concern for child development.

### The microbiome

Although it has been clearly demonstrated that pathogenic microbes can cause brain disorders, there is increasing evidence that the microbial population harboured in the human body, termed the human microbiome, can as a whole influence brain activity<sup>68</sup>. Recent clinical studies among healthy subjects suggest that treatment with a probiotic is associated with reduced symptoms of stress and depression<sup>69</sup>. There is also evidence of associations between the microbiome and neurological diseases, such as multiple sclerosis and autism spectrum disorder (ASD)<sup>70</sup>. In a recent study using a mouse model of ASD, treatment with probiotics alleviated some behavioural symptoms of the disorder<sup>71</sup>. The composition of the human microbiome shows marked differences between countries<sup>72</sup> and comparative research conducted in high-income countries and LMICs could lead to a better understanding of the part played by the human microbiome in brain disorders, and possible treatment of these disorders with factors that favourably alter the microbiome.

treatment. For HIV-associated opportunistic infections such as cryptococcal meningitis, development of therapies that do not rely on intravenous administration is a priority because capacity for intravenous medication is limited in many LMICs, particularly in rural areas. There is also a need for improved access to new assays that detect antiretroviral therapy resistance (for example, the oligonucleotide ligase assay), as these can guide cART treatment. For many parasitic infections, antiparasitic medications are available, but their efficacy in reducing the neurological or cognitive sequelae remains uncertain. For this reason, as noted in the pathogenesis section, development of adjunctive therapies that target prevention or reduction of nervous-system injury is an important research priority.

Development of low-cost, low-toxicity antimicrobials that work against drug-resistant pathogens is a research priority for several infections, including tuberculous meningitis and neonatal sepsis caused by multiresistant Gram-negative infections. Qualitative studies to better understand medical non-compliance, and to develop innovative solutions to reduce non-compliance through newer technologies, such as mobile devices that support medical and public-health practice, are also needed.

Finally, there is a need for multicentre clinical research trials with sufficient sample sizes to provide definitive answers on the efficacy of

specific interventions. For example, where smaller trials had failed, the Cryptococcal Optimal Anti-Retroviral Timing (COAT) trial conducted in Uganda and South Africa successfully determined that deferred initiation of anti-retroviral therapy in individuals with HIV until 5 weeks after treatment of cryptococcal meningitis improved survival<sup>54</sup>. This study finding is likely to change international guidelines.

### Physical, occupational and cognitive rehabilitation

Whereas physical, occupational and cognitive rehabilitation for individuals with sequelae of CNS infections are routine in developed countries, such interventions are limited in LMICs owing to a lack of trained personnel and prohibitive costs. Thus, research on how to build capacity for rehabilitation and how to support it in the context of LMICs is required. Rehabilitation for cognitive impairment can be successfully implemented in the community using locally available resources or in a tertiary institution using advanced methods. In the community, home stimulation, parenting education and support, and provision of financial support or nutritional support for children enrolled in early child development centres have shown some benefit in improving children's cognition<sup>55,56</sup>. Interventions that target both the carer and the child are more effective than those that include either one<sup>56</sup>. In tertiary centres, computer-based cognitive training programmes have proven effective in improving cognition in African children surviving CNS infections<sup>57,58</sup>. These cognitive training programmes can target specific disabilities; however, they are in their early stages and more research is required to determine the most cost-effective implementable and sustainable programmes for LMICs.

### Operations and implementation research

Operationalization and implementation of known effective interventions is another research priority area. Vaccines for *S. pneumoniae* and *N. meningitidis* are highly effective and have been implemented in some LMICs, but these vaccine-preventable infections continue to affect more than 1 million people each year. Thus, in addition to increased investment in the basic science of vaccines, a major research priority is the assessment of methods to support and implement widespread vaccination in LMICs.

Another example is implementation research related to effective prenatal, perinatal and neonatal care, which would decrease neonatal sepsis. Assessment of effective methods for non-physician health workers to provide medical and preventive care to mothers and newborns in LMICs are needed<sup>59</sup>. The recent increase in neonatal infection with multi-drug-resistant Gram-negative organisms in LMICs<sup>24,25</sup> makes prevention of neonatal sepsis an even greater research priority. Rapid diagnostics and treatment interventions that are successful in field trials also require implementation research for successful wide-scale adoption and appropriate use.

### Capacity building

Capacity building within LMICs is key to the successful reduction or elimination of nervous system complications of infection. The Review on research capacity building in this collection addresses this topic in depth (see page S207). Research priorities specifically in the area of infection-related nervous system morbidity include an increase in the number of clinicians and researchers in infectious disease, neuroscience, neurology and mental health<sup>60</sup>, and the dedication of a portion of LMICs health budgets to infectious disease and mental and neurological health<sup>61</sup>. Research training grants and collaborative research between partners in LMICs and high-income countries specifically in the area of infection-related neurocognitive impairment can also help to build human resource capacity. Physical infrastructure is another priority; without space for laboratories, diagnostic equipment or research clinics, surveillance cannot be performed and information to guide interventions to reduce the burden of these infections cannot be generated. With improved human resource capacity and infrastructure, the development of effective screening instruments, prevention

and treatment, and increased government support to address these infections are more likely to be achieved. Box 2 provides examples of capacity building in Peru and Uganda in the area of infections that affect the nervous system. This was enabled by Fogarty International Center grants for collaborative US–LMICs partnerships in these countries.

## CONCLUSIONS

In the past two decades, an increasing body of evidence implicates infection as a cause of substantial nervous system morbidity in high-income countries and LMICs. The burden is particularly high in LMICs, where infections such as HIV (and associated opportunistic infections), HSV, dengue, bacterial and tuberculous meningitis, malaria, neurocysticercosis, STH, schistosomiasis and other infections affect billions of people annually, and cause substantial neurological morbidity in those individuals. The involvement of the nervous system in infections is often first recognized in LMICs where the prevalence of these infections is higher and where new infections often emerge. Research conducted in these countries can contribute to prevention and cure before these infections become globalized. Research is needed in multiple areas to determine the true burden of disease and to develop point-of-care diagnostic assays for diagnosing infection, vaccines and other interventions for preventing infections; to improve our understanding of the pathogenesis of nervous system disease in these infections; to develop better tools for the assessment of neurological, cognitive and mental health impairment; to develop more effective treatments and preventions for nervous system sequelae, to improve the implementation of successful interventions, and to improve rehabilitation for those with long-term neurocognitive or mental health disabilities. Good research studies in these areas, accompanied by equally strong efforts to implement promising technologies and therapies, could substantially decrease the morbidity and mortality of infections affecting the nervous system in LMICs.

1. Fooks, A. R. *et al.* Current status of rabies and prospects for elimination. *Lancet* **384**, 1389–1399 (2014).
2. Labeaud, A. D., Bashir, F. & King, C. H. Measuring the burden of arboviral diseases: the spectrum of morbidity and mortality from four prevalent infections. *Popul. Health Metr.* **9**, 1 (2011).
3. Whitley, R. in *Infections of the Central Nervous System* (eds Whitley, R. J. *et al.*) 137–157 (Lippincott Williams & Wilkins, 2014).
4. McGrath, N., Anderson, N. E., Croxson, M. C. & Powell, K. F. Herpes simplex encephalitis treated with acyclovir: diagnosis and long term outcome. *J. Neurol. Neurosurg. Psych.* **63**, 321–326 (1997).
5. Richter, R. W. & Shimojo, S. Neurologic sequelae of Japanese B encephalitis. *Neurology* **11**, 553–559 (1961).
6. Dutta K, B. A. in *Neuroinflammation and Neurodegeneration* (ed. Toborek Peterson, M) 309–335 (Springer, 2014).
7. Kleinschmidt-DeMasters, B. K. & Gilden, D. H. Varicella-Zoster virus infections of the nervous system: clinical and pathologic correlates. *Arch. Pathol. Lab. Med.* **125**, 770–780 (2001).
8. Insinga, R. P., Itzler, R. F., Pellissier, J. M., Saddier, P. & Nikas, A. A. The incidence of herpes zoster in a United States administrative database. *J. Gen. Internal Med.* **20**, 748–753 (2005).
9. Swanson, E. C. & Schleiss, M. R. Congenital cytomegalovirus infection: new prospects for prevention and therapy. *Pediatr. Clin. North Am.* **60**, 335–349 (2013).
10. Manicklal, S., Emery, V. C., Lazzarotto, T., Boppana, S. B. & Gupta, R. K. The “silent” global burden of congenital cytomegalovirus. *Clin. Microbiol. Rev.* **26**, 86–102 (2013).
11. Bhatt, S. *et al.* The global distribution and burden of dengue. *Nature* **496**, 504–507 (2013).
12. Robin, S. *et al.* Neurologic manifestations of pediatric chikungunya infection. *J. Child Neurol.* **23**, 1028–1035 (2008).
13. World Health Organization. *Number of people (all ages) living with HIV*. [http://www.who.int/gho/hiv/epidemic\\_status/cases\\_all/en/](http://www.who.int/gho/hiv/epidemic_status/cases_all/en/) (2015).
14. Ances, B. M. & Ellis, R. J. Dementia and neurocognitive disorders due to HIV-1 infection. *Semin. Neurol.* **27**, 86–92 (2007).
15. Tan, I. L., Smith, B. R., von Geldern, G., Mateen, F. J. & McArthur, J. C. HIV-associated opportunistic infections of the CNS. *Lancet Neurol.* **11**, 605–617 (2012).
16. Spudich, S. & Meyer, A. C. HIV Neurology. Preface. *Semin. Neurol.* **34**, 5–6 (2014).
17. Jarvis, J. N. & Harrison, T. S. HIV-associated cryptococcal meningitis. *AIDS* **21**, 2119–2129 (2007).
18. Desalerms, A., Kourkoumpetis, T. K. & Mylonakis, E. Update on the epidemiology and management of cryptococcal meningitis. *Expert. Opin. Pharmacother.* **13**, 783–789 (2012).
19. Falusi, O. *et al.* Prevalence and predictors of *Toxoplasma* seropositivity in women with and at risk for human immunodeficiency virus infection. *Clin. Infect. Dis.* **35**, 1414–1417 (2002).

## BOX 2 | CAPACITY BUILDING IN UGANDA AND PERU

### Uganda

The Severe Malaria Research Centre in Uganda is an example of how collaborations between local and foreign scientists, with support from the Fogarty International Center, has built a hub for research excellence. Through a National Institutes of Health (NIH) R21 exploratory research grant in 2004, local scientists developed research capacity by involvement in research studies, and grant and manuscript writing. This has since led to further NIH grants (four R01 grants, a U01 grant, a D43 grant and two R34 grants) as well as multiple grants from other agencies. Ugandan and US faculty are principal investigators on these grants. A book on neuropsychology of African children and more than 30 research papers have been published from these projects so far. Ugandan scientists and physicians have obtained faculty positions in Makerere University, Kampala. The infrastructure that has been built includes high-speed Internet connectivity for research offices and faculty members, a laboratory, a data room and a grants management office. With the infrastructure in place, this centre is now providing training for many Ugandan and US students and researchers at all levels of training, from undergraduates to post-doctoral fellows and faculty.

### Peru

Through the Fogarty International Center NIH R21 and R01 grants in Peru, a network of neurologists who are engaged in brain-disorder research has been developed throughout the country. Trainee alumni of this network now serve as collaborators on emerging research and training activities in both infectious and chronic diseases of the nervous system (such as cerebrovascular diseases). The 2 sites in Lima have been scaled up to 12 hospitals and 2 universities in 3 Peruvian regions. Capacity building of individuals was provided through workshops, hybrid virtual/in-person certificate courses, as well as medium- and long-term training in Seattle and Peru. An initial mentor-training workshop developed into a growing network of mentors, three of whom have been awarded Clayton–Dedonder Mentorship Fellowships by the Fogarty International Center and have started institutionalization of mentor training programmes at three institutions in Lima. Those who received the R21 and R01 grants in the past are now experienced researchers who are leading the development of research in new areas, such as neurogenetics research and the development of a cerebrovascular diseases research training programme. Research supported by these awards resulted in 26 peer-reviewed publications and book chapters. Programme alumni are becoming leaders in brain research and are mentoring the newest wave of young neurologists and neuroscientists.

20. Baud, O. & Aujard, Y. Neonatal bacterial meningitis. *Handb. Clin. Neurol.* **112**, 1109–1113 (2013).
21. Seale, A. C. *et al.* Neonatal severe bacterial infection impairment estimates in South Asia, sub-Saharan Africa, and Latin America for 2010. *Pediatr. Res.* **74** (Suppl.), 73–85 (2013).
22. Gray, K. J., Bennett, S. L., French, N., Phiri, A. J. & Graham, S. M. Invasive group B streptococcal infection in infants, Malawi. *Emerg. Infect. Dis.* **13**, 223–229 (2007).
23. Iregbu, K. C., Elegba, O. Y. & Babaniyi, I. B. Bacteriological profile of neonatal septicaemia in a tertiary hospital in Nigeria. *Afr. Health Sci.* **6**, 151–154 (2006).
24. Kayange, N., Kamugisha, E., Mwizambolya, D. L., Jeremiah, S. & Mshana, S. E. Predictors of positive blood culture and deaths among neonates with suspected neonatal sepsis in a tertiary hospital, Mwanza-Tanzania. *BMC Pediatr.* **10**, 39 (2010).
25. Mehar, V. *et al.* Neonatal sepsis in a tertiary care center in central India: microbiological profile, antimicrobial sensitivity pattern and outcome. *J. Neonatal Perinatal Med.* **6**, 165–172 (2013).
26. van de Beek, D. Progress and challenges in bacterial meningitis. *Lancet* **380**, 1623–1624 (2012).
27. Brouwer, M. C., McIntyre, P., Prasad, K. & van de Beek, D. Corticosteroids for acute bacterial meningitis. *Cochrane Database Syst. Rev.* **6**, CD004405 (2013).
28. Thwaites, G. E., van Toorn, R. & Schoeman, J. Tuberculous meningitis: more questions, still too few answers. *Lancet Neurol.* **12**, 999–1010 (2013).

29. World Health Organization. Global incidence and prevalence of selected curable sexually transmitted infections — 2008. 20 (WHO, 2008).
30. John, C. C. et al. Cerebral malaria in children is associated with long-term cognitive impairment. *Pediatrics* **122**, e92–e99 (2008).
31. Bangirana, P. et al. Severe malarial anemia is associated with long-term neurocognitive impairment. *Clin. Infect. Dis* **59**, 336–344 (2014).
32. Birbeck, G. L. et al. Blantyre Malaria Project Epilepsy Study (BMPES) of neurological outcomes in retinopathy-positive paediatric cerebral malaria survivors: a prospective cohort study. *Lancet Neurol.* **9**, 1173–1181 (2010).
33. Nankabirwa, J. et al. Asymptomatic *Plasmodium* infection and cognition among primary schoolchildren in a high malaria transmission setting in Uganda. *Am. J. Trop. Med. Hyg.* **88**, 1102–1108 (2013).
34. Fleury, A. et al. High prevalence of calcified silent neurocysticercosis in a rural village of Mexico. *Neuroepidemiology* **22**, 139–145 (2003).
35. Carabin, H. et al. Clinical manifestations associated with neurocysticercosis: a systematic review. *PLoS Negl. Trop. Dis.* **5**, e1152 (2011).
36. Ferrari, T. C. & Moreira, P. R. Neuroschistosomiasis: clinical symptoms and pathogenesis. *Lancet Neurol.* **10**, 853–864 (2011).
37. Finkelstein, J. L., Schleinitz, M. D., Carabin, H. & McGarvey, S. T. Decision-model estimation of the age-specific disability weight for schistosomiasis japonica: a systematic review of the literature. *PLoS Negl. Trop. Dis.* **2**, e158 (2008).
38. Coyle, C. M. Schistosomiasis of the nervous system. *Handb. Clin. Neurol.* **114**, 271–281 (2013).
39. Kvalsvig, J. & Albonico, M. Effects of geohelminth infections on neurological development. *Handb. Clin. Neurol.* **114**, 369–379 (2013).
40. Deckers, N. & Dorny, P. Immunodiagnosis of *Taenia solium* taeniosis/cysticercosis. *Trends Parasitol.* **26**, 137–144 (2010).
41. Bang, A. T. et al. Is home-based diagnosis and treatment of neonatal sepsis feasible and effective? Seven years of intervention in the Gadchiroli field trial (1996 to 2003). *J. Perinatol.* **25** (Suppl.), 62–71 (2005).
42. Murray, C. J. et al. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* **380**, 2197–2223 (2012).
43. Alarcon, F., Vanormelingen, K., Moncayo, J. & Vinan, I. Cerebral cysticercosis as a risk factor for stroke in young and middle-aged people. *Stroke* **23**, 1563–1565 (1992).
44. Lengeler, C. Insecticide-treated bed nets and curtains for preventing malaria. *Cochrane Database Syst. Rev.* Cd000363 (2004).
45. Strode, C., Donegan, S., Garner, P., Enayati, A. A. & Hemingway, J. The impact of pyrethroid resistance on the efficacy of insecticide-treated bed nets against African anopheline mosquitoes: systematic review and meta-analysis. *PLoS Med.* **11**, e1001619 (2014).
46. DeBiasi, R. L., Kleinschmidt-DeMasters, B. K., Richardson-Burns, S. & Tyler, K. L. Central nervous system apoptosis in human herpes simplex virus and cytomegalovirus encephalitis. *J. Infect. Dis.* **186**, 1547–1557 (2002).
47. Cobbs, C. S. et al. Human cytomegalovirus infection and expression in human malignant glioma. *Cancer Res.* **62**, 3347–3350 (2002).
48. Peterson, P. K. & Toborek, M. (eds). *Neuroinflammation and Neurodegeneration* (Springer, 2014).
49. Thwaites, G. E. et al. Dexamethasone for the treatment of tuberculous meningitis in adolescents and adults. *N. Engl. J. Med.* **351**, 1741–1751 (2004).
50. Hoffman, S. L. et al. Immunity to malaria and naturally acquired antibodies to the circumsporozoite protein of *Plasmodium falciparum*. *N. Engl. J. Med.* **315**, 601–606 (1986).
51. Good, M. F. et al. Human T-cell recognition of the circumsporozoite protein of *Plasmodium falciparum*: immunodominant T-cell domains map to the polymorphic regions of the molecule. *Proc. Natl Acad. Sci. USA* **85**, 1199–1203 (1988).
52. Stoute, J. A. et al. A preliminary evaluation of a recombinant circumsporozoite protein vaccine against *Plasmodium falciparum* malaria. *N. Engl. J. Med.* **336**, 86–91 (1997).
53. Agnandji, S. T. et al. First results of phase 3 trial of RTS,S/AS01 malaria vaccine in African children. *N. Engl. J. Med.* **365**, 1863–1875 (2011).
54. Boulware, D. R. et al. Timing of antiretroviral therapy after diagnosis of cryptococcal meningitis. *N. Engl. J. Med.* **370**, 2487–2498 (2014).
55. Boivin, M. J. et al. A year-long caregiver training program improves cognition in pre-school Ugandan children with human immunodeficiency virus. *J. Pediatr.* **163**, 1409–1416 (2013).
56. Engle, P. L. et al. Strategies for reducing inequalities and improving developmental outcomes for young children in low-income and middle-income countries. *Lancet* **378**, 1339–1353 (2011).
57. Bangirana, P. et al. Immediate neuropsychological and behavioral benefits of computerized cognitive rehabilitation in Ugandan pediatric cerebral malaria survivors. *J. Dev. Behav. Pediatr.* **30**, 310–318 (2009).
58. Boivin, M. J. et al. A pilot study of the neuropsychological benefits of computerized cognitive rehabilitation in Ugandan children with HIV. *Neuropsychology* **24**, 667–673 (2010).
59. Waiswa, P. et al. The Uganda Newborn Study (UNEST): an effectiveness study on improving newborn health and survival in rural Uganda through a community-based intervention linked to health facilities — study protocol for a cluster randomized controlled trial. *Trials* **13**, 213 (2012).
60. Bruckner, T. A. et al. The mental health workforce gap in low- and middle-income countries: a needs-based approach. *Bull. WHO* **89**, 184–194 (2011).
61. World Health Organization. *Mental Health Atlas 2011*. (WHO, 2011).
62. Liu, L. et al. Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000. *Lancet* **379**, 2151–2161 (2012).
63. Black, R. E. et al. Global, regional, and national causes of child mortality in 2008: a systematic analysis. *Lancet* **375**, 1969–1987 (2010).
64. Walker, C. L. F. et al. Global burden of childhood pneumonia and diarrhoea. *Lancet* **381**, 1405–1416 (2013).
65. Guerrant, R. L., DeBoer, M. D., Moore, S. R., Scharf, R. J. & Lima, A. A. M. The impoverished gut—a triple burden of diarrhoea, stunting and chronic disease. *Nature Rev. Gastroenterol. Hepatol.* **10**, 220–229 (2013).
66. Walker, C. L. F. et al. Does childhood diarrhea influence cognition beyond the diarrhea-stunting pathway? *PLoS ONE* **7**, e47908–e47908 (2012).
67. Richard, S. A. et al. Catch-up growth occurs after diarrhea in early childhood. *J. Nutr.* **144**, 965–971 (2014).
68. Collins, S. M., Surette, M. & Bercik, P. The interplay between the intestinal microbiota and the brain. *Nature Rev. Microbiol.* **10**, 735–742 (2012).
69. Foster, J. A. & McVey Neufeld, K. A. Gut–brain axis: how the microbiome influences anxiety and depression. *Trends Neurosci.* **36**, 305–312 (2013).
70. Ochoa-Reparaz, J., Mielcarz, D. W., Begum-Haque, S. & Kasper, L. H. Gut, bugs, and brain: role of commensal bacteria in the control of central nervous system disease. *Ann. Neurol.* **69**, 240–247 (2011).
71. Hsiao, E. Y. et al. Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. *Cell* **155**, 1451–1463 (2013).
72. Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K. & Knight, R. Diversity, stability and resilience of the human gut microbiota. *Nature* **489**, 220–230 (2012).

# SUPPLEMENTARY INFORMATION

Is linked to the online version of this paper at: <http://dx.doi.org/10.1038/nature16033>

# ACKNOWLEDGEMENTS

The authors' work was supported by the Fogarty International Center and by National Institutes of Health grants R01 NS055349 (C.C.J.), D43 NS078280 (C.C.J.), R25 TW009345 (J.R.Z., C.C.J.), R21 NS077466 (H.C.), R01 NS064901 (H.C.), and R01 NS55627 (J.R.Z.). We are grateful to D. Gilden for his insightful review and suggestions.

# COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests. Financial support for publication has been provided by the Fogarty International Center.

# ADDITIONAL INFORMATION

This work is licensed under the Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0>





REVIEW **OPEN**

# A global perspective on the influence of environmental exposures on the nervous system

Desire Tshala-Katumbay<sup>1,2,3</sup>, Jean-Claude Mwanza<sup>4</sup>, Diane S. Rohlman<sup>5,6</sup>, Gladys Maestre<sup>7</sup> & Reinaldo B. Oriá<sup>8</sup>

Economic transitions in the era of globalization warrant a fresh look at the neurological risks associated with environmental change. These are driven by industrial expansion, transfer and mobility of goods, climate change and population growth. In these contexts, risk of infectious and non-infectious diseases are shared across geographical boundaries. In low- and middle-income countries, the risk of environmentally mediated brain disease is augmented several fold by lack of infrastructure, poor health and safety regulations, and limited measures for environmental protection. Neurological disorders may occur as a result of direct exposure to chemical and/or non-chemical stressors, including but not limited to, ultrafine particulate matters. Individual susceptibilities to exposure-related diseases are modified by genetic, epigenetic and metagenomic factors. The existence of several uniquely exposed populations, including those in the areas surrounding the Niger Delta or north western Amazon oil operations; those working in poorly regulated environments, such as artisanal mining industries; or those, mostly in sub-Saharan Africa, relying on cassava as a staple food, offers invaluable opportunities to advance the current understanding of brain responses to environmental challenges. Increased awareness of the brain disorders that are prevalent in low- and middle-income countries and investments in capacity for further environmental health-related research are positive steps towards improving human health.

*Nature* 527, S187–S192 (19 November 2015), DOI: 10.1038/nature16034

This article has not been written or reviewed by *Nature* editors. *Nature* accepts no responsibility for the accuracy of the information provided.

**R**eports from the World Health Organization (WHO) indicate that the global burden of disease is determined by patterns of disease and disability in low- and middle-income countries (LMICs), which, predictably, have their own environmental signatures ([http://www.who.int/healthinfo/global\\_burden\\_disease/about/en/](http://www.who.int/healthinfo/global_burden_disease/about/en/)). However, the effect of such signatures on both brain health and region or global disability-adjusted life years (DALYs) remains unknown and needs to be added to the agenda of global environmental health research. As for high-income countries, environmental health research programmes in LMICs must primarily focus on elucidating the entire range and source of exposures to define the human ‘exposome’ (the measure of all the exposures of an individual in their lifetime and how these exposures relate to health) relevant to brain health in LMICs. The research agenda should also include mechanistic and translational research, as well as capacity building to foster a new generation of environmental health scientists.

## SCOPE

In this Review, we focus on environmental risk factors for brain diseases and conditions in LMICs (<http://data.worldbank.org/about/country-and-lending-groups>). An iterative search of the literature was conducted using PubMed to retrieve information related to environmental determinants and mechanisms of brain disease in LMICs. Additional opinion was obtained from interviews with leading environmental

scientists and neuroscientists, as well as programme officers at the US National Institutes of Health and US National Institute of Environmental Health Sciences (NIEHS), and Fogarty International Center. This Review integrates the goals and approaches to environmental health research as per the NIEHS 2012–2017 strategic plan (<https://www.niehs.nih.gov/about/strategicplan/>).

## ENVIRONMENTAL EXPOSURE AND BRAIN HEALTH

LMICs are home to around 80–85% of the world’s population<sup>1</sup>. Of these 5.8 billion people<sup>2</sup>, 1 billion remain in extreme poverty, living below the US\$1.25 per day poverty line<sup>3</sup>. Around 3 billion people do not have piped drinking water in their home and 173 million people rely on the direct use of surface water. Without proper sanitation, about one billion continue to defecate in gutters, in the open bush or in open water bodies<sup>4</sup>. Wildfires and deforestation are commonplace and drought and floods, possibly due to climate change, degrade the existing farming systems and create food insecurity<sup>5–7</sup>. Armed conflicts and population displacements impose a toll on human life<sup>8</sup>. Industrial expansion coexists with an unprecedented rise in artisanal mining and unprotected labour<sup>9</sup>. In some instances, normal urbanization operations, such as road construction and quarantines (for example during Ebola outbreaks in the Democratic Republic of the Congo) have created conditions that exacerbated the risk of environmental exposure and brain disease<sup>10</sup>. Flawed regulations compounded by a lack of infrastructure set the stage for

<sup>1</sup>Department of Neurology, Oregon Health & Science University, Portland, Oregon, 97239, USA. <sup>2</sup>National Institute of Biomedical Research, 1197 Kinshasa I, Congo.

<sup>3</sup>Department of Neurology, University of Kinshasa, 825 Kinshasa XI, Congo. <sup>4</sup>Department of Ophthalmology, University of North Carolina at Chapel Hill, North Carolina 27599, USA. <sup>5</sup>Occupational and Environmental Health, The University of Iowa, Iowa 52242, USA. <sup>6</sup>Oregon Institute of Occupational Health Sciences, Oregon Health and Science University, Portland, Oregon, 97239, USA. <sup>7</sup>G. H. Sergievsky Center, Columbia University Medical Center, New York, New York 10032, USA. <sup>8</sup>Department of Morphology and Institute of Biomedicine, Faculty of Medicine, Federal University of Ceara, Fortaleza 60020, Brazil. Correspondence should be addressed to D. T.-K. e-mail: [tshalad@ohsu.edu](mailto:tshalad@ohsu.edu).

**Table 1** | Heavy metals and exposure-related outcomes

Heavy metal	Source of exposure	Susceptibility window	Neurological outcomes	Proposed mechanisms
<b>Lead</b> <sup>76,78</sup>	Lead-contaminated dust, lead-based paint, soil, drinking water, air, leaded gasoline, toys and lead-contaminated sweets	Lifelong	Visual and verbal memory decline, intellectual deficits, decline in executive functioning (fine motor function, hand-eye coordination and reaction time) and hyperactivity in children	Disruption of neurotransmitter release and function, and prenatal disruption of neuronal migration and differentiation; aggravating factors include poor nutrition (deficiency in iron, zinc and calcium) and younger age
<b>Mercury</b> <sup>79,80</sup>	Mining industry, power plants, crematoria, charcoal industry, and contaminated food (mostly sea food) and water	From neural development to neurulation, and adolescence	Ataxia in adults and language, attention, and visuospatial performance deficits in children	Oxidative stress or impairment of intracellular calcium and glutamate homeostasis
<b>Arsenic</b> <sup>80,81</sup>	Contaminated food and drinking water, air and arsenic-based treatments	5–15 years	Impaired selective and focused attention and long-term memory in children, and sensorimotor polyneuropathy	Oxidative stress or disruption of metabolism of neurotransmitters
<b>Copper</b> <sup>82</sup>	Contaminated drinking water and food, uncoated copper cookware and infant formula containing copper	Children Those over 65	Alzheimer's disease, OCD, ADHD, antisocial behaviour and anxiety in children	Oxidative stress, microglia cell activation or promotion of $\alpha$ -synuclein and fibril formation
<b>Cobalt</b> <sup>82,83</sup>	Contaminated drinking water and food, inhalation of dust containing cobalt particles in various industries	Prenatal, young children and the elderly	Optic, auditory and peripheral neuropathy, motor deficits and verbal memory loss	Alteration of mitochondrial oxidative phosphorylation or depletion of neurotransmitters
<b>Cadmium</b> <sup>83</sup>	Fumes or dust, cigarette smoke, and contaminated food and water	Prenatal, young children and the elderly	Antisocial behaviour and attention impairment in children, parkinsonism and peripheral neuropathy	Oxidative damage and neurotransmitter disruption
<b>Manganese</b> <sup>76,79</sup>	Airborne as fumes, aerosols or suspended particulate matter and contaminated water	Childhood and the elderly	Reduced IQ, impaired verbal learning and working and immediate memory in children, and Parkinson-like symptoms	Disruption of mitochondrial respiratory chain reaction; aggravating factors include iron deficiency and impaired biliary excretion (liver injury or disease)
<b>Aluminium</b> <sup>84</sup>	Contaminated air, water and food, cosmetics (such as antiperspirants), metal industries and pharmaceuticals	Lifelong	Alzheimer's pathology in the form of neurofibrillary tangles	Disruption of mitochondrial respiratory chain reaction or inflammation; zinc deficiency acknowledged as an aggravating factor

ADHD, attention-deficit hyperactivity disorder; OCD, obsessive compulsive disorder.

environmental degradation and pollution to pose serious threats — of a chemical or non-chemical nature — to human health. The degradation of local ecosystems leads to the creation of ‘microenvironments’ that have a high risk of harmful exposures, often resulting in unique challenges and increased risk of human disease (Fig. 1).

## HIGH-RISK POPULATIONS AND MICROENVIRONMENTS

Risk of exposure-related brain disease is determined by age, gender and microenvironments created by natural disasters in which economic, social and cultural determinants of health often have important roles. One example of a profit-mediated environmental risk is that caused by the oil industry through accidental spills or mismanagement of oil operations. For instance, crude oil operations have polluted large areas of rainforests, including streams and rivers in Ecuador, Peru and Colombia<sup>11</sup>. The population of Nigeria has faced similar challenges owing to reoccurring oil spills as a result of ageing, ill-maintained or sabotaged pipelines in the Niger Delta. The impact of such man-made and preventable natural disasters on human health has yet to be determined. Effects on human health will depend on the type and composition of the spilled oils, which often contain a mixture of polycyclic hydrocarbons that are known to be toxic to the nervous system<sup>11</sup>. Oil spills arise owing to reasons, such as a lack of vigilance, neglect of necessary health and safety checks, or sometimes even promotion of commercial interests at the expense of communities. Symptoms of acute exposure to raw oil include consistent episodes of headache, nausea, dizziness and fatigue. Chronic effects include psychological disorders, endocrine abnormalities and genotoxic effects<sup>12</sup>.

Microenvironments in which the population has a higher susceptibility to exposure-related diseases have also been created by extreme poverty and natural disasters, including drought and flooding that can degrade soils, plants and farming operations. The burden of conventional neurodevelopmental stressors (for example, lead) on children is exacerbated by unique environmental challenges, including malnutrition and enteric infections<sup>13–16</sup> and, possibly, a diet of neurotoxicant-containing plants such as cassava (*Manihot esculenta*; also known

as tapioca), the grass pea *Lathyrus sativus* or the seeds from the cycad plants, which are all known to be associated with a high burden of neurodisabilities at a population level<sup>17–22</sup>. Populations with unique exposures and risks include those living in the tropical cassava belt of Angola, the Central African Republic, Cameroon, Congo, Tanzania, Uganda, Nigeria and Mozambique<sup>23–30</sup>; those reliant on *L. sativus* as a staple food in Ethiopia, Eritrea, India and Bangladesh<sup>20,31–33</sup>; and the people of the Pacific island Guam or the Japanese Kii Peninsula where the rates of environmentally linked syndromes such as amyotrophic lateral sclerosis-parkinsonism-dementia complex (ALS/PDC) have been declining for reasons that have yet to be uncovered<sup>34,35</sup>.

The impact of early childhood diseases that lead to a vicious cycle of enteric infections and malnutrition has been underestimated and neglected, especially in areas that lack acceptable levels of hygiene and sanitation and that have reduced accessibility to vaccines and antimicrobials. This has caused clinically silent, chronic-illness-related effects, which jeopardize the child's full cognitive development<sup>13,15</sup>. This vicious cycle establishes what is called environmental enteropathy, a mostly subclinical condition (even without diarrhoea) caused by various degrees of intestinal barrier dysfunction, luminal-to-blood intestinal bacterial translocation, low-grade local and systemic inflammation, and disrupted innate intestinal immune responses that may affect growth<sup>36</sup> and cognition<sup>37</sup> and possibly lead to neurodegeneration as well as liver, and metabolic diseases later in life<sup>38,39</sup>.

Adolescents in LMICs experience a higher burden of exposures (in contrast with those in high-income countries), primarily because of the childhood labour crisis. Although there are regulations and international agreements restricting child labour, often there are exceptions for certain industries, notably the growing agricultural industry, one of the most hazardous industries worldwide<sup>40,41</sup>. In this context, adolescent workers are at risk of exposure to agrochemicals such as pesticides<sup>42,43</sup>. Other work-related threats include exposure to organic solvents in work that involves painting and manufacture, to toxic metals and fine particulate matters in artisanal mining, and to heat and ambient air pollution while working long hours and outside<sup>41</sup>. Exposure to

Table 2 | Organic compounds and exposure-related outcomes

Organic compound	Source of exposure	Susceptibility window	Neurological outcomes	Proposed mechanisms
<b>Bisphenol A</b> <sup>85,86</sup>	Food from cans with linings that contain BPA, and contaminated food and water	Prenatal and childhood	Anxious behaviour, hyperactivity and depressive behaviour, and learning impairment in children	Unclear, but females seem to be more susceptible
<b>Phthalates</b> <sup>86–88</sup>	Food or drink that has been in contact with containers or products containing phthalates, and air and dust containing phthalates	Prenatal and childhood	Depressive and conduct-related behaviours (ODD, attention problems, rule-breaking and aggressive behaviour in children)	Oxidative stress
<b>Organophosphates</b> <sup>89,90</sup>	Contaminated food and water, polluted air and professional dermal contact	Lifelong	Neurodevelopmental deficits, impaired attention and working memory, impaired speed and executive functions, and delayed peripheral polyneuropathy	Inhibition of acetyl-cholinesterase
<b>Organochlorinated compounds (DDT/PCBs)</b> <sup>91</sup>	Contaminated food, drinking water and air	Prenatal and lifelong	Impaired intellectual ability, ADHD-like behaviours and locomotor deficits	Disruption of neurotransmitter function, oxidative stress or derangement of calcium homeostasis; children seem to be more susceptible.
<b>Organobromide compounds (PBDEs)</b> <sup>92,93</sup>	Contaminated food, water and air	Lifelong	IQ deficits, impaired attention, fine motor coordination and cognition functioning in children	Impairment of thyroid hormone homeostasis
<b>Organic solvents</b> <sup>94,95</sup>	Air and professional dermal contact and glue sniffing	Adolescence and adulthood	Headache, memory deficits, and central and peripheral neuropathy	Protein adduction and/or oxidative misfolding or oxidative stress
<b>Food-born neurotoxins (cassava cyanogenic glucosides and BOAA in <i>Lathyrus sativus</i>)<sup>96–98</sup> or contaminants (fungal toxins)</b>	Oral ingestion	Lifelong	Spastic paraparesis, cognition deficits and possibly convulsive disorders	Oxidative stress, excitotoxicity and protein carbamylation for cassava cyanogens; children and females seem to be more susceptible; malnutrition is acknowledged as an aggravating factor

ADHD, attention-deficit hyperactivity disorder; BOAA, beta-(N)-oxalyl-amino-L-alanine acid; BPA, bisphenol A; DDT/PCBs, dichlorodiphenyltrichloroethane/polychlorinated biphenols; ODD, oppositional defiant disorder.

Table 3 | Complex exposures and neurological outcomes

Exposure	Source of exposure	Susceptibility window	Neurological outcomes	Proposed mechanisms
<b>Coal/charcoal burning</b> <sup>99,100</sup>	Charcoal/coal combustion, gas grilling, wood smoke, or coal mine dust or ash	Lifelong	Neurological signs of exposure to arsenic	Oxidative stress
<b>Car emissions</b> <sup>101,102</sup>	Contaminated air	Lifelong	Learning disability and motor impairment	Oxidative stress or neurotransmission disruption
<b>Fine and ultrafine particulate matters</b> <sup>103</sup>	Air pollution from car or construction equipment exhausts, wood burning, heating oil or coal, forest fires, volcanic eruptions, tobacco smoke and cooking	Lifelong	Behavioural and decreased IQ, impaired fluid cognition, memory and executive functions, and possibly autism	Oxidative stress, neurotransmission disruption or neuroinflammation

industrial solvents such as *n*-hexane, for example, may occur because of poor safety regulations or recreational glue sniffing. This may result in headache, acute encephalopathy or sensorimotor neuropathies that are reversible on cessation<sup>44</sup>.

Adults in LMICs may be at a particularly high risk of environmental exposure and related brain diseases compared with those in high-income countries. In general, they experience a higher burden of disease owing to a lifetime of cumulative exposures and co-morbidities that are highly prevalent in LMICs. The latter include malaria, nutritional deficiencies and neurotropic infections such as those caused by human T-cell lymphotropic viruses (HTLV). For example, it was reported that endemic foci of HTLV-I-associated myelopathy coexist with outbreaks of konzo (a spastic paraparesis linked to the toxicity of cassava cyanogens) in some regions of the Congo<sup>45,46</sup>. In these areas, women of child-bearing age are particularly susceptible to the toxicity of cassava cyanogens for reasons that have not been elucidated, although they may be linked to hormonal influences and poor nutrition<sup>47</sup>.

## PATHWAYS TO BRAIN DISEASE

Exposure-related brain damage may result from chemical and/or non-chemical stressors. Damage to the nervous system often leads to a range of bilateral and symmetrical motor and/or sensory symptoms. Behavioural problems, cognition deficits and psychiatric illness may also occur. Non-chemical stressors include, but are not limited to, psychological stress, heat, noise, fine and ultrafine particulate matter (FUPM), and waterborne, airborne or foodborne pathogens that may

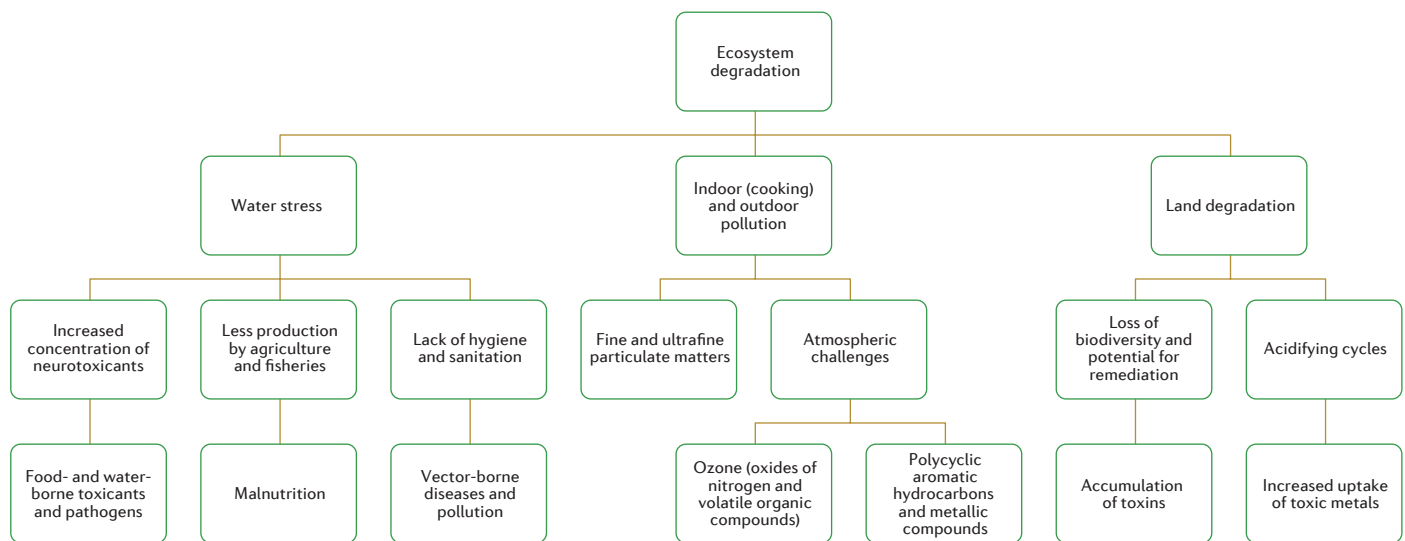
occur under the conceptual framework shown in Fig. 1. Chemicals with neurotoxic potential that people are commonly exposed to are listed in Tables 1–3. Mixed exposure, for example chemical-covered FUPM from industrial emissions; co-exposure to chemical and non-chemical stressors; and repeated and multiple exposure can occur, creating a complex human environmental exposome.

Brain damage linked to chemical exposure may result from chemicals interfering with neurotransmission through molecular mimicry or reacting with crucial biomolecules and causing incorrect function (for example, protein or DNA adduction and/or crosslinking). For both chemical and non-chemical exposures, the mechanisms of brain damage may include injury to the vascular system (for example, fine particulate matter induced vascular pathology), systemic dyshomeostasis (for example, cadmium-induced kidney disease) and hormonal imbalance (for example, through endocrine disruption; Table 1).

The susceptibility to exposure-related disease is, however, determined by mechanisms of functional genetics, epigenetics and metagenomics at the interface between risk factors and neurological outcomes (Fig. 2).

It is increasingly acknowledged that genetic and epigenetic factors, including the effect of maternal stress on brain function, influence the effect of environmental exposure<sup>48,49</sup>. For example, the E4 allele of the APOE gene that is reportedly associated with higher risk of late-onset Alzheimer's disease, although not in people from sub-Saharan Africa and with a mild association among Hispanic people, is associated with protection against early childhood diarrhoea and its related cognitive





**Figure 1** Environmental (chemical and non-chemical) threats to brain health in low- and middle-income countries. Multiple sources of exposure (air, water and food) coexist, and malnutrition and vector-borne diseases, notably infections, compound the risk of brain disease. Co-exposures not shown include heat, psychological stress and a poor physical environment, such as crowding.

impairment<sup>50–52</sup>. One example of gene–environment interactions is the relationship between air pollution components and the gene encoding the MET receptor tyrosine kinase. Several studies have implicated *MET* as an autism risk gene<sup>53–55</sup>. Stratification of the risk conferred by a functional promoter variant in this gene (rs1858830) and by local traffic-related air pollution (regional particulate matter less than 10 micrometres in diameter and nitrogen dioxide exposure) revealed significant multiplicative interaction between the risk genotype and the air pollution exposure<sup>56</sup>.

Our knowledge of the pathways that lead to late onset of exposure-related neurological disease is still sparse<sup>57,58</sup>. Many studies suggest that the genetic and environmental causes of late onset diseases act in parallel and share common molecular mechanisms<sup>59</sup>. A number of studies have supported the concept that early-life exposure to pollutants reprograms global gene expression in old age through epigenetic mechanisms<sup>60–63</sup>. Variation in exposure response, even among individuals exposed to the same environment could be due not only to early-life exposures, but also to differences in genetic make up<sup>64–66</sup>. The extent and nature of exposures and related brain diseases in LMICs provide opportunities to explore and overcome the long reach that childhood exposure has into adulthood, as well as provide us with new advances in environmental health sciences<sup>67</sup>.

Exposure-related neurological deficits in LMICs range from peripheral neuropathies to a large number of acute, subacute or chronic central nervous system diseases. Deficits may occur prenatally, or during childhood or adolescence, and may be carried through to old age. Clinical implications include, but are not limited to, neural tube defects, learning disabilities, behavioural problems, psychiatric disorders, cognitive decline and the occurrence of distinct entities such as neurolathyrism, tropical ataxic neuropathy, ALS/PDC and konzo<sup>20,30,35,68,69</sup> (Fig. 3).

The human microbiome may be of particular interest to the mechanistic understanding of exposure-related diseases in LMICs because it may influence the burden of heavy metals<sup>70</sup>, the metabolism of food-borne neurotoxins such as cassava cyanogens<sup>18</sup>, and the outcome of enteric diseases in early life, including the child's neurodevelopmental potential<sup>71–73</sup>.

## RESEARCH AND CAPACITY BUILDING

Recent advances in environmental health sciences have elucidated the myriad risk factors and mechanisms of brain damage that are

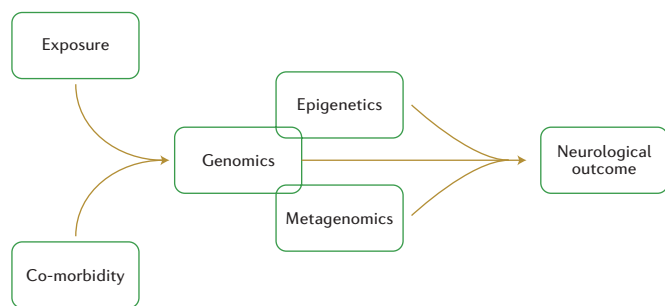
associated with environmental exposures. The existence of uniquely exposed populations in LMICs offers invaluable opportunities to advance our current understanding of brain responses to environmental threats. In some instances, well-characterized neurotoxins may be used as chemical probes to dissect the pathophysiology of the nervous system. However, challenges at the population level still remain, including setting exposure limits and developing metrics and methodologies to assess the long-term impact of environmental exposures on disease burden in LMICs and, therefore, globally. Climate change and mining of rare elements, which may include radioactive materials, present unpredictable risks, and should be added to the environmental health research agenda. The toll of such exposures on the global burden of disease may be efficiently addressed only through competent partnerships and alliances established on a global scale and focused on key areas and priorities (Box 1). Although there is evidence that some of these are already in place, more research and research capacity is needed to continue this agenda to improve human health, globally.

## ONE HEALTH–GLOBAL HEALTH DIMENSIONS

Environmental degradation and contamination, changes in climate and ecosystems, and vector-borne pathogens or neurotoxins are the primary environmental threats to human life and intellectual performance.

### BOX 1 | INTERWOVEN RESEARCH AREAS AND INVESTMENT PRIORITIES IN GLOBAL ENVIRONMENTAL RESEARCH

- Epidemiology and statistical modelling for exposure and risk assessment in co-exposure and co-morbidity scenarios
- High-throughput ‘-omic’ methodologies
- Bioinformatics and knowledge management
- Development of diagnostic and remediation tools — validation and implementation of environmental sensors, detectors and biomarkers of exposure and related outcomes
- Development of metrics and methodologies to assess the long-term impact of environmental exposures on neurological disease burden
- Understanding the pathways that lead to late onset of exposure-related neurological diseases
- Training and capacity building in the areas listed above



**Figure 2** | Environmental framework and pathways to environmentally induced neurological disease in low- and middle-income countries. Susceptibility to neurological disease is determined at the interface between a particular exposure, epigenetic and metagenetic make up, and the presence of co-morbidities.

Humans, plants and animals adapt to environmental challenges, but some may overcome their adaptive capabilities and create imminent risks for all<sup>17,74</sup>. Strategies to promote human health will therefore require a serious commitment to trans-disciplinary work, plant and animal health and building capacity on a global scale<sup>75</sup>.

1. Sumner, A. *Global Poverty and the New Bottom Billion: What if Three-Quarters of the World's Poor Live in Middle-Income Countries?* 1–43 (IDS, 2010).
2. World Bank. Data. *Low and Middle Income Countries* <http://data.worldbank.org/income-level/LMY> (World Bank, 2015).
3. World Bank. *Global Monitoring Report 2014/2015: Ending Poverty and Sharing Prosperity* (World Bank, 2015).
4. Lang, V. & Lingnau, H. Defining and measuring poverty and inequality post-2015. *J. Int. Develop.* **27**, 399–414 (2015).
5. Haile, M. Weather patterns, food security and humanitarian response in sub-Saharan Africa. *Phil. Trans. R. Soc. Lond. B* **360**, 2169–2182 (2005).
6. Bjorklund, G. Workshop 4 (synthesis): securing food production under climate variability — exploring the options. *Water Sci. Technol* **49**, 147–149 (2004).
7. Kim, K. H., Kabir, E. & Ara Jahan, S. A review of the consequences of global climate change on human health. *J. Environ. Sci. Health C. Environ. Carcinog. Ecotoxicol. Rev.* **32**, 299–318 (2014).
8. Rieder, M. & Choonara, I. Armed conflict and child health. *Arch. Dis. Child.* **97**, 59–62 (2012).
9. Seccatore, J. et al. An estimation of the artisanal small-scale production of gold in the world. *Sci. Total Environ.* **496**, 662–667 (2014).
10. Banea, M., Tylleskar, T. & Rosling, H. Konzo and ebola in Bandundu region of Zaire. *Lancet* **349**, 621 (1997).
11. Jernelov, A. The threats from oil spills: now, then, and in the future. *Ambio* **39**, 353–366 (2010).
12. Levy, B. S. & Nassetta, W. J. The adverse health effects of oil spills: a review of the literature and a framework for medically evaluating exposed individuals. *Int. J. Occup. Environ. Health* **17**, 161–167 (2011).
13. Guerrant, R. L. et al. The impoverished gut — a triple burden of diarrhoea, stunting and chronic disease. *Nature Rev. Gastroenterol. Hepatol.* **10**, 220–229 (2013).
14. Guerrant, R. L. et al. Magnitude and impact of diarrheal diseases. *Arch. Med. Res.* **33**, 351–355 (2002).
15. Guerrant, R. L. et al. Malnutrition as an enteric infectious disease with long-term effects on child development. *Nutr. Rev.* **66**, 487–505 (2008).
16. Petri, W. A. Jr. et al. Enteric infections, diarrhea, and their impact on function and development. *J. Clin. Invest.* **118**, 1277–1290 (2008).
17. Wang, W. et al. Cassava genome from a wild ancestor to cultivated varieties. *Nature Commun.* **5**, 5110 (2014).
18. Tshala-Katumbay, D. et al. Cassava food toxins, konzo disease, and neurodegeneration in sub-Saharan Africans. *Neurology* **80**, 949–951 (2013).
19. Sarmiento, A. et al. Valorization of traditional foods: nutritional and bioactive properties of *Cicer arietinum* L. and *Lathyrus sativus* L. pulses. *J. Sci. Food Agric.* **95**, 179–185 (2015).
20. Spencer, P. S. & Schaumburg, H. H. Lathyrism: a neurotoxic disease. *Neurobehav. Toxicol. Teratol.* **5**, 625–629 (1983).
21. Marler, T. E. & Lindstrom, A. J. Free sugar profile in cycads. *Front. Plant. Sci.* **5**, 526 (2014).
22. Kisby, G. E. & Spencer, P. S. Is neurodegenerative disease a long-latency response to early-life genotoxin exposure? *Int. J. Environ. Res. Public Health* **8**, 3889–3921 (2011).
23. Banea, J. P. et al. Effectiveness of wetting method for control of konzo and reduction of cyanide poisoning by removal of cyanogens from cassava flour. *Food Nutr. Bull.* **35**, 28–32 (2014).
24. Tylleskar, T. et al. Konzo in the Central African Republic. *Neurology* **44**, 959–961 (1994).
25. Ciglenecki, I. et al. Konzo outbreak among refugees from Central African Republic in Eastern region, Cameroon. *Food Chem. Toxicol.* **49**, 579–582 (2011).



**Figure 3** | Neurocognition deficits in konzo, a disease linked to eating cyanogenic cassava. **a**, Spasticity in a 14-year old boy severely affected by konzo. **b**, Deficits in mental processing are evident from the results of a neuropsychological test.

26. Nzwalo, H. & Cliff, J. Konzo: from poverty, cassava, and cyanogen intake to toxic-nutritional neurological disease. *PLoS Negl. Trop. Dis.* **5**, e1051 (2011).
27. Mlingi, N. L. et al. Recurrence of konzo in southern Tanzania: rehabilitation and prevention using the wetting method. *Food Chem. Toxicol.* **49**, 673–677 (2011).
28. Cliff, J. et al. Konzo associated with war in Mozambique. *Trop. Med. Int. Health* **2**, 1068–1074 (1997).
29. Okitundu Luwa, E. A. D. et al. Persistence of konzo epidemics in Kahemba, Democratic Republic of Congo: phenomenological and socio-economic aspects. *Pan. Afr. Med. J.* **18**, 213 (2014).
30. Oluwole, O. S. et al. Persistence of tropical ataxic neuropathy in a Nigerian community. *J. Neurol. Neurosurg. Psych.* **69**, 96–101 (2000).
31. Tekle-Haimanot, R. et al. Clinical and electroencephalographic characteristics of epilepsy in rural Ethiopia: a community-based study. *Epilepsy Res.* **7**, 230–239 (1990).
32. Ludolph, A. C. et al. Studies on the aetiology and pathogenesis of motor neuron diseases. 1. Lathyrism: clinical findings in established cases. *Brain* **110**, 149–165 (1987).
33. Ngudi, D. D. et al. Research on motor neuron diseases konzo and neuropathology: trends from 1990 to 2010. *PLoS Negl. Trop. Dis.* **6**, e1759 (2012).
34. Lee, S. E. Guam dementia syndrome revisited in 2011. *Curr. Opin. Neurol.* **24**, 517–524 (2011).
35. Kaji, R. et al. ALS-parkinsonism-dementia complex of Kii and other related diseases in Japan. *Parkinsonism Relat. Disord.* **18** (Suppl 1), S190–S191 (2012).
36. Prendergast, A. J. et al. Stunting is characterized by chronic inflammation in Zimbabwean infants. *PLoS ONE* **9**, e86928 (2014).
37. Patrick, P. D. et al. Limitations in verbal fluency following heavy burdens of early childhood diarrhea in Brazilian shantytown children. *Child Neuropsychol.* **11**, 233–244 (2005).
38. Korpe, P. S. & Petri, W. A. Jr. Environmental enteropathy: critical implications of a poorly understood condition. *Trends Mol. Med.* **18**, 328–336 (2012).
39. Petri, W. A., Naylor, C. & Haque, R. Environmental enteropathy and malnutrition: do we know enough to intervene? *BMC Med.* **12**, 187 (2014).
40. Tilman, D. & Clark, M. Global diets link environmental sustainability and human health. *Nature* **515**, 518–522 (2014).
41. Ferguson, K. T. et al. The physical environment and child development: an international review. *Int. J. Psychol.* **48**, 437–468 (2013).
42. Crane, A. L. et al. Longitudinal assessment of chlorpyrifos exposure and effect biomarkers in adolescent Egyptian agricultural workers. *J. Expo. Sci. Environ. Epidemiol.* **23**, 356–362 (2013).
43. Rohlman, D. S. et al. Characterizing exposures and neurobehavioral performance in Egyptian adolescent pesticide applicators. *Metab. Brain Dis.* **29**, 845–855 (2014).
44. Spencer, P. S. et al. The enlarging view of hexacarbon neurotoxicity. *Crit. Rev. Toxicol.* **7**, 279–356 (1980).
45. Tylleskar, T. et al. Konzo, an epidemic spastic paraparesis in Africa, is not associated with antibodies to HTLV-I, HIV, or HIV-gag-encoded proteins. *J. Acquir. Immune Defic. Syndr. Hum. Retroviro.* **12**, 317–318 (1996).
46. Jeannel, D. et al. The risk of tropical spastic paraparesis differs according to ethnic group among HTLV-I carriers in Inongo, Zaire. *J. Acquir. Immune Defic. Syndr.* **6**, 840–844 (1993).
47. Tylleskar, T. et al. Dietary determinants of a non-progressive spastic paraparesis (Konzo): a case-referent study in a high incidence area of Zaire. *Int. J. Epidemiol.* **24**, 949–956 (1995).
48. Vidal, A. C. et al. Maternal stress, preterm birth, and DNA methylation at imprint regulatory sequences in humans. *Genet. Epigenet.* **6**, 37–44 (2014).
49. Bale, T. L. Lifetime stress experience: transgenerational epigenetics and germ cell programming. *Dialogues Clin. Neurosci.* **16**, 297–305 (2014).
50. Maestre, G. et al. Apolipoprotein E and Alzheimer's disease: ethnic variation in genotypic risks. *Ann. Neurol.* **37**, 254–259 (1995).
51. Oria, R. B. et al. ApoE polymorphisms and diarrheal outcomes in Brazilian shantytown children. *Braz. J. Med. Biol. Res.* **43**, 249–256 (2010).
52. Oria, R. B. et al. APOE4 protects the cognitive development in children with heavy diarrhea burdens in Northeast Brazil. *Pediatr. Res.* **57**, 310–316 (2005).
53. Jackson, P. B. et al. Further evidence that the rs1858830 C variant in the promoter region of the MET gene is associated with autistic disorder. *Autism Res.* **2**, 232–236 (2009).

54. Sousa, I. et al. MET and autism susceptibility: family and case-control studies. *Eur. J. Hum. Genet.* **17**, 749–758 (2009).
55. Peng, Y. et al. MET receptor tyrosine kinase as an autism genetic risk factor. *Int. Rev. Neurobiol.* **113**, 135–165 (2013).
56. Volk, H. E. et al. Autism spectrum disorder: interaction of air pollution with the MET receptor tyrosine kinase gene. *Epidemiology* **25**, 44–47 (2014).
57. Charleta, L. et al. Neurodegenerative diseases and exposure to environmental metals Mn, Pb, and Hg. *Coord. Chem. Rev.* **256**, 2147–2163 (2012).
58. Oteiza, P. I., Mackenzie, G. G. & Verstraeten, S. V. Metals in neurodegeneration: involvement of oxidants and oxidant-sensitive transcription factors. *Mol. Aspects Med.* **25**, 103–115 (2004).
59. Ali, S. F., Binienda, Z. K. & Imam, S. Z. Molecular aspects of dopaminergic neurodegeneration: gene-environment interaction in parkin dysfunction. *Int. J. Environ. Res. Public Health* **8**, 4702–4713 (2011).
60. Dosunmu, R., Alashwal, H. & Zawia, N. H. Genome-wide expression and methylation profiling in the aged rodent brain due to early-life Pb exposure and its relevance to aging. *Mech. Ageing Dev.* **133**, 435–443 (2012).
61. Bihagi, S. W. et al. Infantile postnatal exposure to lead (Pb) enhances tau expression in the cerebral cortex of aged mice: relevance to AD. *Neurotoxicology* **44**, 114–120 (2014).
62. Wang, G. et al. Early life origins of metabolic syndrome: the role of environmental toxicants. *Curr. Environ. Health Rep.* **1**, 78–89 (2014).
63. Collotta, M., Bertazzi, P. A. & Bollati, V. Epigenetics and pesticides. *Toxicology* **307**, 35–41 (2013).
64. Singh, S. et al. Influence of CYP2C9, GSTM1, GSTT1 and NAT2 genetic polymorphisms on DNA damage in workers occupationally exposed to organophosphate pesticides. *Mutat. Res.* **741**, 101–108 (2012).
65. Morahan, J. M. et al. Genetic susceptibility to environmental toxicants in ALS. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **144B**, 885–890 (2007).
66. Goodrich, J. M. & Basu, N. Variants of glutathione s-transferase pi 1 exhibit differential enzymatic activity and inhibition by heavy metals. *Toxicol. In Vitro* **26**, 630–635 (2012).
67. Currie, E. & Vogl, T. Early-life health and adult circumstance in developing countries. *Ann. Rev. Econom.* **5**, 1–36 (2013).
68. Tshala-Katumbay, D. et al. Analysis of motor pathway involvement in konzo using transcranial electrical and magnetic stimulation. *Muscle Nerve* **25**, 230–235 (2002).
69. Boivin, M. J. et al. Neuropsychological effects of konzo: a neuromotor disease associated with poorly processed cassava. *Pediatrics* **131**, e1231–e1239 (2013).
70. Breton, J. et al. Gut microbiota limits heavy metals burden caused by chronic oral exposure. *Toxicol. Lett.* **222**, 132–138 (2013).
71. Alonso, C. et al. Intestinal barrier function and the brain-gut axis. *Adv. Exp. Med. Biol.* **817**, 73–113 (2014).
72. Sommer, F. & Backhed, F. The gut microbiota — masters of host development and physiology. *Nature Rev. Microbiol.* **11**, 227–238 (2013).
73. Arrieta, M. C. et al. The intestinal microbiome in early life: health and disease. *Front. Immunol.* **5**, 427 (2014).
74. Gleadow, R. M. et al. Growth and nutritive value of cassava (*Manihot esculenta* Cranz.) are reduced when grown in elevated CO<sub>2</sub>. *Plant Biol.* **11** (Suppl 1), 76–82 (2009).
75. Erisman, J. W. et al. Put people at the centre of global risk management. *Nature* **519**, 151–153 (2015).
76. Nean, A. & Guillarte, T. Mechanisms of heavy metal neurotoxicity: lead and manganese. *Toxicol. Res.* **2**, 99–114 (2013).
77. Mason, L. H., Harp, J. P. & Han, D. Y. Pb neurotoxicity: neuropsychological effects of lead toxicity. *Biomed. Res. Int.* **2014**, 840547 (2014).
78. Sanders, T. et al. Neurotoxic effects and biomarkers of lead exposure: a review. *Rev. Environ. Health* **24**, 15–45 (2009).
79. Farina, M., Rocha, J. B. & Aschner, M. Mechanisms of methylmercury-induced neurotoxicity: evidence from experimental studies. *Life Sci.* **89**, 555–563 (2011).
80. Ercal, N., Gurer-Orhan, H. & Aykin-Burns, N. Toxic metals and oxidative stress part I: mechanisms involved in metal-induced oxidative damage. *Curr. Top. Med. Chem.* **1**, 529–539 (2001).
81. Florea, A. M. & Busselberg, D. Occurrence, use and potential toxic effects of metals and metal compounds. *Biomaterials* **19**, 419–427 (2006).
82. Valko, M., Morris, H. & Cronin, M. T. Metals, toxicity and oxidative stress. *Curr. Med. Chem.* **12**, 1161–1208 (2005).
83. Catalani, S. et al. Neurotoxicity of cobalt. *Hum. Exp. Toxicol.* **31**, 421–437 (2012).
84. Kumar, V. & Gill, K. D. Aluminium neurotoxicity: neurobehavioural and oxidative aspects. *Arch. Toxicol.* **83**, 965–978 (2009).
85. Harley, K. G. et al. Prenatal and early childhood bisphenol A concentrations and behavior in school-aged children. *Environ. Res.* **126**, 43–50 (2013).
86. Yoltos, K. et al. Prenatal exposure to bisphenol A and phthalates and infant neurobehavior. *Neurotoxicol. Teratol.* **33**, 558–566 (2011).
87. Engel, S. M. et al. Prenatal phthalate exposure is associated with childhood behavior and executive functioning. *Environ. Health Perspect.* **118**, 565–571 (2010).
88. Miodovnik, A. et al. Endocrine disruptors and childhood social impairment. *Neurotoxicology* **32**, 261–267 (2011).
89. Jamal, G. A. et al. A clinical neurological, neurophysiological, and neuropsychological study of sheep farmers and dippers exposed to organophosphate pesticides. *Occup. Environ. Med.* **59**, 434–441 (2002).
90. Blanc-Lapierre, A. et al. Cognitive disorders and occupational exposure to organophosphates: results from the PHYTONER study. *Am. J. Epidemiol.* **177**, 1086–1096 (2013).
91. Fonnum, F. & Mariussen, E. Mechanisms involved in the neurotoxic effects of environmental toxicants such as polychlorinated biphenyls and brominated flame retardants. *J. Neurochem.* **111**, 1327–1347 (2009).
92. Costa, L. G. et al. A mechanistic view of polybrominated diphenyl esters developmental neurotoxicity. *Toxicol. Lett.* **15**, 282–294 (2014).
93. Linares, V., Belles, M. & Domingo, J. L. Human exposure to PBDE and critical evaluation of health hazards. *Arch. Toxicol.* **89**, 335–356 (2015).
94. Viaene, M. Overview of the neurotoxicants effects in solvent-exposed workers. *Arch. Public Health* **60**, 217–232 (2002).
95. Tshala-Katumbay, D. et al. New insights into mechanisms of gamma-diketone-induced axonopathy. *Neurochem. Res.* **34**, 1919–1923 (2009).
96. Kassa, R. M. et al. On the biomarkers and mechanisms of konzo, a distinct upper motor neuron disease associated with food (cassava) cyanogenic exposure. *Food Chem. Toxicol.* **49**, 571–578 (2011).
97. Spencer, P. S. Food toxins, ampa receptors, and motor neuron diseases. *Drug Metab. Rev.* **31**, 561–587 (1999).
98. Makila-Mabe, B. G. et al. Serum 8,12-iso-iPF<sub>2</sub>α-VI isoprostane marker of oxidative damage and cognition deficits in children with konzo. *PLoS ONE* **9**, e107191 (2014).
99. Kang, Y. et al. Arsenic in Chinese coals: distribution, modes of occurrence, and environmental effects. *Sci. Total Environ.* **412–413**, 1–13 (2011).
100. Liu, J. et al. Chronic arsenic poisoning from burning high-arsenic-containing coal in Guizhou, China. *Environ. Health Perspect.* **110**, 119–122 (2002).
101. Block, M. L. et al. Nanometer size diesel exhaust particles are selectively toxic to dopaminergic neurons: the role of microglia, phagocytosis, and NADPH oxidase. *FASEB J.* **18**, 1618–1620 (2004).
102. Kilburn, K. H. Effects of diesel exhaust on neurobehavioral and pulmonary functions. *Arch. Environ. Health* **55**, 11–17 (2000).
103. Costa, L. G. et al. Neurotoxicants are in the air: convergence of human, animal, and in vitro studies on the effects of air pollution on the brain. *Biomed. Res. Int.* **2014**, 736385 (2014).

#### ACKNOWLEDGEMENTS

All the authors are thankful to NIEHS and Fogarty International Centre for research grant support and the scientific expertise of their respective programme officers A. Kirshner and K. Michels and staff members. The intellectual contribution of R. Kalaria of Newcastle University, UK, is very much appreciated.

#### COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests. Financial support for publication has been provided by the Fogarty International Center.

#### ADDITIONAL INFORMATION



This work is licensed under the Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0>



REVIEW **OPEN**

# Global neurotrauma research challenges and opportunities

Andrés M. Rubiano<sup>1</sup>, Nancy Carney<sup>2</sup>, Randall Chesnut<sup>3</sup> & Juan Carlos Puyana<sup>4</sup>

Traumatic injury to the brain or spinal cord is one of the most serious public health problems worldwide. The devastating impact of 'trauma', a term used to define the global burden of disease related to all injuries, is the leading cause of loss of human potential across the globe, especially in low- and middle-income countries. Enormous challenges must be met to significantly advance neurotrauma research around the world, specifically in underserved and austere environments. Neurotrauma research at the global level needs to be contextualized: different regions have their own needs and obstacles. Interventions that are not considered a priority in some regions could be a priority for others. The introduction of inexpensive and innovative interventions, including mobile technologies and e-health applications, focused on policy management improvement are essential and should be applicable to the needs of the local environment. The simple transfer of a clinical question from resource-rich environments to those of low- and middle-income countries that lack sophisticated interventions may not be the best strategy to address these countries' needs. Emphasis on promoting the design of true 'ecological' studies that include the evaluation of human factors in relation to the process of care, analytical descriptions of health systems, and how leadership is best applied in medical communities and society as a whole will become crucial.

*Nature* 527, S193–S197 (19 November 2015), DOI: 10.1038/nature16035

This article has not been written or reviewed by *Nature* editors. *Nature* accepts no responsibility for the accuracy of the information provided.

**T**he global burden of disease (GBD) related to all injuries or 'trauma' is the leading cause of loss of human potential around the world especially in low- and middle-income countries (LMICs). According to the 2010 report of the GBD study 89% of trauma-related deaths occur in LMICs. Nearly 6 million people die each year as a result of trauma, accounting for 10% of the world's deaths — 32% more than the number of fatalities from malaria, tuberculosis and HIV/AIDS combined<sup>1</sup>.

Within the spectrum of trauma-related injuries, traumatic brain injury (TBI) and spinal cord injury (SCI) are the largest causes of death and disability, leading to suffering by, and costs to, the individual, their family and society. Social costs include changes to the family care structure owing to cognitive, emotional and/or physical disabilities in addition to economic costs and reduced productivity. The incidence of central nervous system (CNS) injuries varies between regions, with estimates ranging from 200 to 600 injuries per 100,000 people. The data are sparse and the true incidence of both TBI and SCI may be considerably underestimated<sup>2</sup>.

Efforts to quantify the magnitude of TBI are hindered by several factors, the most common of which is related to the lack of consistent data recording where this occurs<sup>3–5</sup>. For example, the absence of formal injury surveillance or reporting systems (trauma registries) in some high-income countries as well as in LMICs, leads to an underestimate of the true magnitude of CNS burden of disease worldwide. Despite 89% of the trauma population being in LMICs, pre-hospital mortality for CNS injuries is not systematically recorded in research

that originates in these countries. Even fewer LMICs have formally implemented a data-specific registry for neurotrauma. In addition, most patients with TBI have mild to moderate injury and are therefore often not reported<sup>6,7</sup>. CNS injuries in patients with multiple trauma, especially as a result of military or civilian conflicts, may be recorded under other causes of death or injury statistical codes.

## EPIDEMIOLOGY AND GLOBAL RESEARCH IN TBI

Although high-quality worldwide data of TBI incidence and prevalence are difficult to find, neurotrauma registries from high-income countries indicate that around 5.3 million people in the United States and nearly 7.7 million people in Europe are living with TBI-related disability. The 2010 GBD study<sup>8</sup> shows that in high-income countries an important cause of TBI is motor vehicle accidents, and that there has been a shift in the age of the affected population towards older groups. In LMICs, those with TBI are generally young adult pedestrians, cyclists or motorcyclists. In regions where the prevalence of armed violence is higher (Central America, the Middle East and Central Africa), assault and gunshot injuries are important causes of TBI<sup>9</sup>. Deficits associated with TBI, including impaired attention, poor executive function, depression, impulsivity, poor decision-making and aggressive behaviour, have particularly striking social and economic consequences for individuals, families and the development of societies as a whole<sup>10,11</sup>. An example of the heterogeneity of the data in international epidemiological research in TBI is shown in Supplementary Table 1.

<sup>1</sup>Neuroscience Institute, Neurotrauma Group, El Bosque University, Avenue Carrera 9a, 131A-02, Edificio Fundadores, Bogotá, Colombia. <sup>2</sup>Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, 3181 SW Sam Jackson Park Road, Portland, Oregon 97239, USA. <sup>3</sup>Neurological Surgery Global Health, Harborview Medical Center, Department of Neurosurgery, 325 Ninth Avenue, Seattle, Washington 98104, USA. <sup>4</sup>Surgery Critical Care Medicine and Clinical Translational Sciences Director of Global Surgery, University of Pittsburgh, 200 Lothrop Street, Pittsburgh, Pennsylvania 15213, USA. Correspondence should be addressed to J. C. P. e-mail: puyanajc@upmc.edu.

Basic and clinical research in TBI have been focused on understanding the biological process of the disease, developing advanced diagnostic tools, minimizing secondary brain injury and improving treatment guidelines. Unfortunately, the evidence generated from neurotrauma studies carried out in high-income countries does not always translate to LMICs, where the health infrastructure (including providers and facilities) is limited, creating a different context for care practice<sup>12–14</sup>. Recently published consensus statements, established in high-income countries, do not take into account the unique challenges that neurotrauma researchers may face in LMICs. Most of these evidence-based recommendations are best applied in well-funded and well-equipped neurosurgical or neurotrauma centres<sup>15–17</sup>. The recent Benchmark Evidence from South American Trials: Treatment of Intracranial Pressure (BEST Trip) trial was based on standard recommendations for randomized clinical trials in high-income countries<sup>18–22</sup>. Results were far from expected because advanced monitoring tools used to guide treatments in high-income countries were not as successful in LMICs, and discussion within the global neurotrauma scientific community emerged after the publication of this study<sup>23–27</sup>. The interpretation and implications of the study for the neurotrauma field in high-income countries and LMICs are still being analysed<sup>28</sup>. The applicability of high-income-country clinical research standards in LMICs is an important topic for future international research.

Research will focus on new trends for TBI care, including, but not limited to, the use of hyperosmolar fluids, blood components for early resuscitation and other strategies aimed at improving resuscitation in patients with multiple injuries, including TBI<sup>29–31</sup>. Additional aspects that are more applicable to LMICs, such as the importance of data collection (neurotrauma registries), capacity building for advanced education in neurotrauma-care provision and research, and integration of teams within a trauma-care system have been recently proposed<sup>32,33</sup>. Treatment strategies such as prophylactic hypothermia have also been considered as therapies with the potential for further research in LMICs. However, this would require organizational effort by the health-care systems to obtain reliable evidence. Multicentre collaborative approaches towards data collection are already in place in high-income countries; such endeavours may also be an efficient and productive strategy for TBI research in LMICs. Alcohol and substance misuse associated with TBI is a further key topic that needs to be addressed in LMICs. Poorer outcomes have been associated with alcohol and substance misuse in high-income countries, but the findings were mixed and further research is required in different contexts<sup>34</sup>.

## EPIDEMIOLOGY AND GLOBAL RESEARCH IN SCI

In 2013, the World Health Organization (WHO) and the International Spinal Cord Society (ISCoS) joined together to report SCIs worldwide<sup>35,36</sup>. Unsurprisingly, similar to TBI, information in epidemiological studies from LMICs was limited. Since 2000, at least 7 papers have reviewed the epidemiology of SCI around the world<sup>37–43</sup> and 2 papers have focused on the epidemiology of SCI in LMICs<sup>44,45</sup>. Common conclusions relate to the lack of information available in LMICs owing to the absence of SCI registries — paradoxically, these are regions where incidence of the disease is high according to observational studies. Reported incidence ranged from 12 to close to 60 cases per million inhabitants in different countries (see Supplementary Table 2). It is difficult to compare the data owing to the heterogeneity of the studies, which had diverse methods for reporting and classifying the disease.

SCI registries from high-income countries and meta-analysis of studies reporting incidence of the disease allow us to estimate that worldwide, every year nearly 250,000 to 500,000 people sustain an SCI<sup>46,47</sup>. Historically, around 90% of SCIs have been associated with trauma; however, in an analysis of high-income-country registries, non-traumatic SCI has recently increased to beyond 10%. The traumatic SCI population is young, especially in LMICs. Underestimation of SCIs is common and with the exception of a few countries that have countrywide registries (Finland, Canada and the United States),

incidence estimates are extrapolated from city or regional data that may not be representative of the countries as a whole<sup>45</sup>.

The three most common causes of SCI across the world are road traffic accidents, falls and violence. Because of the low quality of data, especially in LMICs, there may be country-level variation in the causes or the context of injury, especially in areas with higher levels of social violence (Central America, the Middle East and Central Africa). In studies from Africa, transportation-related events account for nearly 70% of cases; in countries affected by war such as Afghanistan, around 60% of all SCI cases are related to violence. It has been estimated that work-related injuries contribute to at least 15% of all traumatic SCIs. There are consistently higher incidence rates in adult males, and the two groups most likely to have an SCI are young and elderly males. Life expectancy for patients with an SCI is shorter than the average in LMICs, as well as in comparison with patients with an SCI from high-income countries<sup>48</sup>.

In a similar way to TBIs, most SCI research is carried out in high-income countries, and focused on the basic science of the biological process of the disease, helping to develop treatment guidelines and the application of advanced technology for nerve reconstruction, sophisticated prosthesis and advanced rehabilitation. Many of these studies are not feasible in LMICs where basic science resources are scarce and advanced rehabilitation is almost non-existent. Recommendations by researchers from high-income countries for designing SCI clinical trials have been published, but the applicability of these recommendations to the LMIC context has yet to be determined<sup>49</sup>. Organization of neurotrauma-care systems and capacity building for neurotrauma and SCI registries may have an effect in LMICs, but they are not priorities for researchers from high-income countries. Crucial aspects, such as the relationship between pre-hospital care and outcomes for patients with SCI are difficult to analyse in LMICs because pre-hospital care is not widely available<sup>50</sup>. An example of the difficulties in SCI research owing to a lack of data is presented in a review about pressure ulcers as a complication in patients with SCI in LMICs<sup>51</sup>. Over the past few years, researchers in China have been making important steps towards evaluating cellular therapies and improving the quality of life of those with SCI. Registries are now available to improve epidemiological data collection within the country<sup>52</sup>.

## Lessons learned from clinical neurotrauma research

In this section, we present examples of active neurotrauma research groups from LMICs.

**Latin America.** The three examples from Latin America draw upon our direct experience of working in Argentina, Bolivia, Colombia and Ecuador.

Between 2008 and 2011 a randomized controlled trial of intracranial-pressure (ICP) monitoring in patients with TBI, which compared the management of patients with severe TBI that was based on information from ICP monitoring with treatment that was based on imaging and clinical examination without ICP information, was performed in Bolivia and Ecuador. The study reported no difference in outcomes between these groups. The study is considered to be class 1 — it has high internal validity. Thus, for LMICs, the study provided concrete information on which to base resource-allocation decisions, and documented the clinical success of a treatment approach that is sustainable in low-resource environments. Sufficiently skilled clinical staff with a better organized protocol of care could produce good recovery results in the intensive care unit (ICU) without data from an ICP monitor by using clinical assessment to manage intracranial hypertension<sup>53–56</sup>.

A multicentre randomized controlled trial of post-discharge care of paediatric traumatic brain injury in Argentina aimed to develop, introduce and test a family-provided, post-discharge intervention for children with complicated mild, moderate and severe TBI. Multiple research methods were used, beginning with focus groups with children who had sustained a TBI, as well as with their parents, physicians,

nurses and social workers. The focus group experience was one of the most important aspects of this project. It gave the high-income-country research team an appreciation of the realities of TBI in these communities, and allowed for an ecologically relevant approach to developing an intervention. The participating hospitals elected to maintain the intervention, and said that the protocol improved overall quality of care for the children and their families. An unexpected finding was that despite reports that paediatric TBI is a serious problem in Latin America<sup>57</sup>, hospitals in this study saw, on average, fewer than one child with TBI per month.

In another study, a standardized trauma-care protocol decreased in-hospital mortality of patients with severe TBI in a LMIC teaching hospital<sup>58</sup>. The standardized trauma-care protocol was based on generally accepted best practices; damage-control resuscitation strategies were proposed based on military protocols from war scenarios in the Middle East. With the knowledge that most hospitals in LMICs have financial or logistical limitations in building evidence-based protocols and do not have a pre-existing trauma registry, an administrative electronic database was adapted to capture clinical information about adults with TBI<sup>59</sup>. Adherence to the protocol was limited – around 60%. Surprisingly, the barriers to adherence were not associated with resources or technology, but with human factors related to changing established practices. How to create motivational interventions to change practice is an important research question for LMICs.

**China.** Here we summarize studies of the use of decompressive craniectomy and hypothermia in the treatment of severe TBI conducted in China.

In a study of decompressive craniectomy, the influence of a standard, larger, unilateral frontotemporoparietal bone flap (a standard trauma craniectomy) was compared with a limited, smaller temporo-parietal bone flap (a limited craniectomy) based on a 6-month Glasgow Outcome Scale (GOS) score and complications. The investigators found significantly greater mortality in patients with a limited craniectomy<sup>60</sup>. A second study in China compared 1-month mortality, complications and the 1-year GOS score of larger unilateral decompressive craniectomy with routine unilateral temporo-parietal decompressive craniectomy. One-month survival and one-year GOS scores were significantly better in the larger unilateral decompressive craniectomy group; however, they had a higher rate of complications<sup>61</sup>. These findings contribute important information for LMIC environments where decompressive craniectomy may be the only treatment option available for a quick resolution of ICP.

Similar to decompressive craniectomy, prophylactic hypothermia for the treatment of ICP is a relatively low-technology option available in LMICs. However, its influence on patient outcome is yet to be clearly demonstrated. In a comparison of a longer course of mild hypothermia (33–35°C for 3–14 days) with normothermia, mortality was found to be lower and 1-year GOS scores were better for the hypothermia group<sup>62</sup>. A subsequent study that compared short-term (1–3 days) and long-term (4–6 days) mild hypothermia found that patients given a course of hypothermia for 5 days had significantly better 6-month GOS scores than those given a 2-day course<sup>63</sup>. Finally, patients who received systemic cooling (full body) were compared with those who received selective brain cooling (head only) and normothermia. Pneumonia rates were lowest and 2-year GOS scores were highest in patients who received selective cooling<sup>64</sup>. Although the results of these studies are promising, the findings are tempered by contradictory findings from similar studies conducted in other countries. What is important for research in LMICs is the question of how to manage a crucial aspect of TBI, ICP, in the absence of technological resources.

**India and Nepal.** Recent studies from India and Nepal describe the experience of creating surveillance and research infrastructures in extremely austere conditions.

In India, the WHO's Standards for Surveillance of Neurotrauma

were used to design and build a simple data collection instrument, and an observational study of TBI was conducted in a rural teaching hospital. Over 6 months, data on 414 patients were collected and descriptive statistics about a sample were reported. Logistical difficulties, including a lack of closely managed data collection and entry, inconsistent coding and missing data were recorded<sup>65</sup>. In an epidemiological study of trauma in a hospital in the Eastern region of Nepal, data on 6,793 patients over 1 year were collected, and a subset of TBI cases was reported on. This is the first study in Nepal that collected comprehensive patient profiles and reported outcomes in detail<sup>66</sup>. The authors concluded that trauma-related injury significantly contributes to morbidity and mortality and is the third leading cause of death in the region.

The examples from Latin America, China, and India and Nepal illustrate the vast differences in the spectrum of neurotrauma research across LMICs. Neurotrauma research at the global level needs to be contextualized – different regions have their own needs and challenges during the research process. Certain interventions may be high priority in one country, but low priority in another.

## Research priorities, opportunities and challenges

Although CNS injury is important, we must acknowledge that an isolated brain or spinal cord injury represents a small fraction of the burden of trauma as a disease, but they occur frequently in the context of the multiple-injury patient. From a mechanistic viewpoint, isolated CNS trauma is the best model to understand the pathophysiology of brain or spinal injury, but it is naive to ignore the fact that patients with multiple trauma injuries have a conglomerate of systemic events that affect the brain. We must, therefore, study CNS injury in the setting in which it most commonly occurs: the patient with multiple injuries. This context needs to be part of the research portfolio in global health, especially in LMICs where it is difficult to measure the interactions of different interventions in the same patient. Assessment could, however, improve with better organization of the existing resources.

The impact of new resuscitation techniques, early use of blood products and early evaluation of coagulopathy in patients with TBI or SCI could be key in areas where violence is an important cause of injury and transport times to hospital are long. The impact of non-invasive intracerebral blood detectors, advanced airway management by non-physicians, and pre-hospital resuscitation fluids could be a priority in areas where organized trauma systems do not exist. Most LMICs do not have organized pre-hospital care. If it exists, it is not consistent and there are no evidence-based transport protocols. There is little training for ambulance staff, which may or may not include a physician. If physicians are present, often they are not trained in emergency medicine. It is possible that in LMICs the most important area of research is within the public health system in order to demonstrate the benefits of an organized pre-hospital care system to improve patient outcomes, and to reduce costs both in hospital and post-discharge. Establishing systematic surveillance systems to accurately identify incidence, prevalence, processes of care and outcomes following TBI and SCI are essential priorities for research in areas where these systems do not exist.

## Capacity-building priorities and opportunities

Countries where TBI and SCI are a significant burden of disease are also those with substantial gaps in services that affect the entire spectrum of trauma care, including prevention, pre-hospital care, specialized neurotrauma care, rehabilitation, quality control and research. As daunting as it may seem to propose capacity-building activities in all these areas, the comprehensive management of TBI and SCI will require human resources, infrastructure and research training aimed at enhancing capacity in all these components. Because resources are limited, the next fundamental question is how to establish priorities so that these areas can advance in parallel. Research training grants and collaborative research between partners in LMICs and high-income countries should include the creation of multidisciplinary teams of



health-care professionals that work in prevention, pre-hospital care, clinical care and clinical epidemiology to expand overall human-resource capacity. At the same time, basic training in provision of care is gravely needed, as is research training. The capacity-building curriculum for different parts of the health system will differ, but they must all introduce the concepts that are inherent to the creation of a comprehensive trauma system.

Specific capacity-building activities should address the design and feasibility of using modular theme-based trainee curricula that employ enabling technologies such as e-learning and teleconferencing; low-tech clinical simulation training that emphasizes early life-saving interventions or procedures; and team-training techniques to accentuate the collaborative nature of neurotrauma care. Strengthened capacity to use information and communication technologies to support research and research-training programmes is also needed.

## CONCLUSION

Enormous challenges must be met to significantly advance neurotrauma research worldwide, particularly in underserved areas and austere environments. Experts beyond clinical practitioners and basic science researchers will need to participate in order to meet these challenges. The introduction of inexpensive and innovative interventions, including communication technologies, mobile-health applications and policy management approaches that meet the needs of a particular local environment is the ultimate goal. Simply transferring a clinical question from a resource-rich environment to that of a LMIC which lacks sophisticated interventions may not be the best strategy to address the needs of LMICs. Furthermore, the findings of studies conducted in resource-rich environments may not necessarily result in evidence-based guidelines that can be implemented in health-care scenarios with more-limited resources.

A new context for capacity building in neurotrauma should include broad international collaborations and global-health opportunities directed at creating not only advanced researchers, but also health leaders who work in field research and health-policy development and implementation. Fundamental questions in research that are relevant to LMICs need to go beyond health-care facilities and medical schools. Emphasis on promoting the design of true 'ecological' studies that include evaluation of human factors in relation to the process of care, analytical descriptions of health systems, and how leadership is applied in the medical community and society as a whole will be crucial.

- Norton, R. & Kobusingye, O. Injuries. *N. Engl. J. Med.* **368**, 1723–1730 (2013).
- Reilly, P. The impact of neurotrauma on society: an international perspective. *Prog. Brain Res.* **161**, 3–9 (2007).
- Puvanachandra, P. & Hyder, A. Traumatic brain injury in Latin America and the Caribbean: a call for research. *Salud Pública de México* **50**, S3–S5 (2008).
- Puvanachandra, P. & Hyder, A. The burden of traumatic brain injury in Asia: a call for research. *Pak. J. Neurol. Sci.* **4**, 27–32 (2009).
- Furlan, J. C., Sakakibara, B. M., Miller, W. C. & Krassioukov, A. V. Global incidence and prevalence of traumatic spinal cord injury. *Can. J. Neurol. Sci.* **40**, 456–464 (2013).
- Cassidy, J. D. et al. Incidence, risk factors and prevention of mild traumatic brain injury: results of the WHO Collaborating Centre Task Force on Mild Traumatic Brain Injury. *J. Rehabil. Med.* **43** (Suppl.), 28–60 (2004).
- Centers for Disease Control and Prevention. *Traumatic Brain Injury in the United States: Epidemiology and Rehabilitation, Congress Report 2014* <http://www.biausa.org/announcements/cdc-s-report-to-congress-on-tbi-epidemiology-and-rehabilitation> (CDC, 2014).
- Horton, R. GBD 2010: understanding disease, injury, and risk. *Lancet* **380**, 2053–2054 (2012).
- Rozenbeek, B., Maas, A. I. & Menon, D. K. Changing patterns in the epidemiology of traumatic brain injury. *Nature Rev. Neurol.* **9**, 231–236 (2013).
- Hofman, K., Primack, A., Keusch, G. & Hrynkow S. Addressing the growing burden of trauma and injury in low- and middle-income countries. *Am. J. Public Health* **95**, 13–17 (2005).
- Langlois, J. A., Rutland Brown, W. & Wald, M. M. The epidemiology and impact of traumatic brain injury. A brief overview. *J. Head Trauma Rehabil.* **21**, 375–378 (2006).
- Gosselin, R. A. The increasing burden of injuries in developing countries. Direct and indirect consequences. *Tech. Orthop.* **24**, 230–232 (2009).
- Borse, N. N. & Hyder, A. A. Call for more research on injury from developing world: results of a bibliometric analysis. *Indian J. Med. Res.* **129**, 321–326 (2009).
- Sitsapesan, H. A., Lawrence, T. P., Sweasey, C. & Wester, K. Neurotrauma outside the high-income setting: a review of audit and data collection strategies. *World Neurosurg.* **79**, 568–575 (2013).
- Rubiano, A. M. & Rios, A. M. Neurotrauma research in Latin America. *J. Res. Fund. Care Online* **6**, 1–2 (2014).
- Razmkon, A. Priorities and concerns for research on neurotrauma in the developing world. *Bull. Emerg. Trauma* **1**, 5–6 (2013).
- Rubiano, A. M. Strengthening neurotrauma care in the Pan American Region. *J. Trauma Crit. Care Emerg. Surg.* **2**, 5–6 (2013).
- Thurmond, V. A. et al. Advancing integrated research in psychological health and traumatic brain injury: common data elements. *Arch. Phys. Med. Rehabil.* **91**, 1633–1636 (2010).
- Maas, A. I. et al. Reorientation of clinical research in traumatic brain injury: report of an international workshop on comparative effectiveness research. *J. Neurotrauma* **29**, 32–46 (2012).
- Chesnut, R. M. et al. A trial of intracranial-pressure monitoring in traumatic brain injury. *N. Engl. J. Med.* **367**, 2471–2481 (2012).
- Narayan, R. K. et al. Clinical trials in head injury. *J. Neurotrauma* **19**, 503–557 (2002).
- Maas, A. I., Roozenbeek, B. & Manley, G. T. Clinical trials in traumatic brain injury: past experience and current developments. *Neurotherapeutics* **7**, 115–126 (2010).
- Chesnut, R. M. et al. Traumatic brain injury in Latin America: lifespan analysis randomized control trial protocol. *Neurosurgery* **71**, 1055–1063 (2012).
- Rubiano, A. M. & Puyana, J. C. Intracranial pressure monitoring in traumatic brain injury. *N. Engl. J. Med.* **368**, 1748 (2013).
- Le Roux, P. Intracranial pressure after the BEST Trip trial: a call for more monitoring. *Curr. Opin. Crit. Care* **20**, 141–147 (2014).
- Mattei, T. Intracranial pressure monitoring in severe traumatic brain injury: who is still bold enough to keep sinning against level I evidence? *World Neurosurg.* **79**, 602–604 (2013).
- Sahuquillo, J. & Biestro, A. Is intracranial pressure monitoring still required in the management of severe traumatic brain injury? Ethical and methodological considerations on conducting clinical research in poor and low income countries. *Surg. Neurol. Int.* **5**, 86 (2014).
- Tosetti, P. et al. Toward an international initiative for traumatic brain injury research. *J. Neurotrauma* **30**, 1211–1222 (2013).
- Yue, J. K. et al. Transforming research and clinical knowledge in traumatic brain injury pilot: multicenter implementation of the common data elements for traumatic brain injury. *J. Neurotrauma* **30**, 1831–1844 (2013).
- Maas, A. I. et al. Advancing care for traumatic brain injury: findings from the impact studies and perspectives on future research. *Lancet Neurol.* **12**, 1200–1210 (2013).
- Green, S. E. et al. Improving the care of people with traumatic brain injury through the Neurotrauma Evidence Translation (NET) program: protocol for a program of research. *Implement. Sci.* **7**, 74 (2012).
- Bayley, M. T. et al. Where to build the bridge between evidence and practice? Results of an international workshop to prioritize knowledge translation activities in traumatic brain injury care. *J. Head Trauma Rehabil.* **29**, 268–276 (2014).
- Rubiano, A. M., Puyana, J. C., Mock, C. N., Bullock, M. R. & Adelson, P. D. Strengthening neurotrauma care systems in low and middle-income countries. *Brain Injury* **27**, 262–272 (2013).
- Parry-Jones, B. L., Vaughan, F. L. & Cox, W. M. Traumatic brain injury and substance misuse: a systematic review of prevalence and outcomes research (1994–2004). *Neuropsych. Rehab.* **16**, 537–560 (2006).
- Sorensen, F. B. et al. IPSCI: a WHO and ISCOS collaboration report. *Spinal Cord* **52**, 87 (2014).
- Bickembach, J. *International Perspectives on Spinal Cord Injury* (WHO/International Spinal Cord Society, 2013).
- Wyndaele, M. & Wyndaele, J. J. Incidence, prevalence and epidemiology of spinal cord injury: what learns a worldwide literature survey? *Spinal Cord* **44**, 523–529 (2006).
- Ackery, A., Tator, C. & Krassioukov, A. A global perspective on spinal cord injury epidemiology. *J. Neurotrauma* **21**, 1355–1370 (2004).
- Cripps, R. A. et al. A global map for traumatic spinal cord injury epidemiology: towards a living data repository for injury prevention. *Spinal Cord* **49**, 493–501 (2011).
- Van der Berg, M. E., Castellote, J. M., Mahillo-Fernandez, I. & de Pedro-Cuesta, J. Incidence of spinal cord injury worldwide: a systematic review. *Neuroepidemiology* **34**, 184–192 (2010).
- Furlan, J. C., Sakakibara, B. M., Miller, W. C. & Krassioukov, A. V. Global incidence and prevalence of traumatic spinal cord injury. *Can. J. Neurol. Sci.* **40**, 456–464 (2013).
- Singh, A., Tetraault, L., Kalsy-Ryan, S., Nouri, A. & Fehlings, M. G. Global prevalence and incidence of traumatic spinal cord injury. *Clin. Epidemiol.* **6**, 309–331 (2014).
- Lee, B., Cripps, R. A., Fitzharris, M. & Wing, P. C. The global map for traumatic spinal cord injury epidemiology: update 2011, global incidence rate. *Spinal Cord* **52**, 110–116 (2014).
- Chiu, W. T. et al. Epidemiology of traumatic spinal cord injury: comparisons between developed and developing countries. *Asia Pac. J. Public Health* **22**, 9–18 (2010).

45. Rahimi-Vovaghar, V. et al. Epidemiology of traumatic spinal cord injury in developing countries: a systematic review. *Neuroepidemiology* **41**, 65–85 (2013).
46. Furlan, J. C. Databases and registries on traumatic spinal cord injury in Canada. *Can. J. Neurol. Sci.* **40**, 454–455 (2013).
47. National Spinal Cord Injury Statistical Center. *The 2013 Annual Statistical Report for the Spinal Cord Injury Model Systems* (Univ. Alabama at Birmingham, 2013).
48. Oderud, T. Surviving spinal cord injury in low-income countries. *African J. Disability* **3**, 1–9 (2014).
49. Tuszyński, M. H. et al. Guidelines for the conduct of clinical trials for spinal cord injury as developed by the ICCP Panel: clinical trial inclusion/exclusion criteria and ethics. *Spinal Cord* **45**, 222–231 (2007).
50. Nielsen, K. et al. Assessment of status of prehospital care in 13 low and middle-income countries. *Prehosp. Emerg. Care* **16**, 381–389 (2012).
51. Zaczek, E. C., Creasey, G. & Crew, J. D. Pressure ulcers in people with spinal cord injury in developing nations. *Spinal Cord* **53**, 7–13 (2015).
52. Li, J. et al. The epidemiological survey of acute traumatic spinal cord injury (ATSCI) of 2002 in Beijing municipality. *Spinal Cord* **49**, 777–782 (2011).
53. Chesnut, R. M., Petroni, G. & Rondina, C. Intracranial-pressure monitoring in traumatic brain injury. *N. Engl. J. Med.* **368**, 1751–1752 (2013).
54. Petroni, G. et al. Early prognosis of severe traumatic brain injury in a urban Argentinian trauma center. *J. Trauma* **68**, 564–570 (2010).
55. Ghajar, J. & Carney, N. Intracranial-pressure monitoring in traumatic brain injury. *N. Engl. J. Med.* **368**, 1749 (2013).
56. Sarrafzadeh, A. S., Smoll, N. R. & Unterberg, A. W. Lessons from the intracranial pressure-monitoring trial in patients with traumatic brain injury. *World Neurosurg.* **82**, 393–395 (2014).
57. Murgio, A. et al. Minor head injury at paediatric age in Argentina. *J. Neurosurg. Sci.* **43**, 15–23 (1999).
58. Kesinger M. R., Puyana J. C. & Rubiano A. M. Improving trauma care in low-and middle-income countries by implementing a standardized trauma protocol. *World J. Surg.* **38**, 1869–1874 (2014).
59. Kesinger, M. R. et al. A standardized trauma care protocol decreased in hospital mortality of patients with severe traumatic brain injury at a teaching hospital in a middle-income country. *Injury* **45**, 1350–1354 (2014).
60. Jiang, J. Y. et al. Efficacy of standard trauma craniectomy for refractory intracranial hypertension with severe traumatic brain injury: a multicenter, prospective, randomized controlled study. *J. Neurotrauma* **22**, 623–628 (2005).
61. Qiu, W. et al. Effects of unilateral decompressive craniectomy on patients with unilateral acute post-traumatic brain swelling after severe traumatic brain injury. *Crit. Care* **13**, R185 (2009).
62. Jiang, J., Yu, M. & Zhu, C. Effect of long-term mild hypothermia therapy in patients with severe traumatic brain injury: 1-year follow-up review of 87 cases. *J. Neurosurg.* **93**, 546–549 (2000).
63. Jiang, J. Y. et al. Effect of long-term mild hypothermia or short-term mild hypothermia on outcome of patients with severe traumatic brain injury. *J. Cereb. Blood Flow Metab.* **26**, 771–776 (2006).
64. Liu, W. G. et al. Effects of selective brain cooling in patients with severe traumatic brain injury: a preliminary study. *J. Int. Med. Res.* **34**, 58–64 (2006).
65. Agrawal, A. et al. Developing traumatic brain injury data bank: prospective study to understand the pattern of documentation and presentation. *Indian J. Neurotrauma* **9**, 87 (2012).
66. Bajracharya, A., Agrawal, A., Yam, B., Agrawal, C. & Lewis O. Spectrum of surgical trauma and associated head injuries at a university hospital in eastern Nepal. *J. Neurosci. Rural Pract.* **1**, 2 (2010).

#### SUPPLEMENTARY INFORMATION

Is linked to the online version of this paper at: <http://dx.doi.org/10.1038/nature16035>

#### ACKNOWLEDGEMENTS

The authors acknowledge support from MEDITECH Foundation Research Group and South Colombian University Public Health Research Group (J. Montenegro, M. N. Suarez and D. Charry) in the preparation of the tables and references. The authors are supported by NIH R21TW009332-01A1, R25TW009714-01 and R01NS080648-01 grants. We are grateful to J. Povlishock for his insightful review and suggestions.

#### COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests. Financial support for publication has been provided by the Fogarty International Center.

#### ADDITIONAL INFORMATION



This work is licensed under the Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0>

REVIEW **OPEN**

# Regional research priorities in brain and nervous system disorders

Vijayalakshmi Ravindranath<sup>1</sup>, Hoang-Minh Dang<sup>2</sup>, Rodolfo G. Goya<sup>3</sup>, Hader Mansour<sup>4,5</sup>, Vishwajit L. Nimgaonkar<sup>6</sup>, Vivienne Ann Russell<sup>7</sup> & Yu Xin<sup>8</sup>

The characteristics of neurological, psychiatric, developmental and substance-use disorders in low- and middle-income countries are unique and the burden that they have will be different from country to country. Many of the differences are explained by the wide variation in population demographics and size, poverty, conflict, culture, land area and quality, and genetics. Neurological, psychiatric, developmental and substance-use disorders that result from, or are worsened by, a lack of adequate nutrition and infectious disease still afflict much of sub-Saharan Africa, although disorders related to increasing longevity, such as stroke, are on the rise. In the Middle East and North Africa, major depressive disorders and post-traumatic stress disorder are a primary concern because of the conflict-ridden environment. Consanguinity is a serious concern that leads to the high prevalence of recessive disorders in the Middle East and North Africa and possibly other regions. The burden of these disorders in Latin American and Asian countries largely surrounds stroke and vascular disease, dementia and lifestyle factors that are influenced by genetics. Although much knowledge has been gained over the past 10 years, the epidemiology of the conditions in low- and middle-income countries still needs more research. Prevention and treatments could be better informed with more longitudinal studies of risk factors. Challenges and opportunities for ameliorating nervous-system disorders can benefit from both local and regional research collaborations. The lack of resources and infrastructure for health-care and related research, both in terms of personnel and equipment, along with the stigma associated with the physical or behavioural manifestations of some disorders have hampered progress in understanding the disease burden and improving brain health. Individual countries, and regions within countries, have specific needs in terms of research priorities.

*Nature* 527, S198–S206 (19 November 2015), DOI: 10.1038/nature/16036

This article has not been written or reviewed by *Nature* editors. *Nature* accepts no responsibility for the accuracy of the information provided.

As outlined in the introduction to this series (see page S151), the proportion of the global burden of disease (GBD) due to neurological, mental health, developmental and substance-use (NMDs) disorders is rising worldwide<sup>1</sup>. The type of disorder and reason for increase varies across countries<sup>2</sup>, regions and populations as indicated by the regional differences in disability adjusted life years (DALYs; a metric developed to take both mortality and morbidity measures into account). DALYs for a disease or health condition are calculated as the sum of the years of life lost (YLL) due to premature mortality in the population and the years lost due to disability (YLD) for people living with the health condition or its consequences ([http://www.who.int/healthinfo/global\\_burden\\_disease/metrics\\_daly](http://www.who.int/healthinfo/global_burden_disease/metrics_daly)). The first regional use of DALYs, the regional patterns of disability-free life expectancy and disability-adjusted life expectancy, were reported by the Global Burden of Disease Study<sup>3</sup>.

Opportunities to ameliorate nervous system disorders could be increased by both local and regional research collaborations. Lessons learned locally, and those learned in collaboration across regions and

countries, may be adapted and applied to other areas, there may also be opportunities to leverage resources. Some disorders have physical boundaries, whereas others have sociocultural and economic contexts. Thus, the challenges faced in high-income countries are often different from those in low- or middle-income countries (LMICs) in type, characteristic or scale. Population demographics, genetics, income, religion, culture, language, ethnic origin, conflicts, land area and quality, and population size vary widely between and within LMICs. Although there is some commonality in the prevalence of certain brain disorders (Fig. 1), significant diversity exists with respect to the origin, manifestation and treatment strategies or options adopted across these regions. In this Review, we focus on sub-Saharan Africa, the Middle East and North Africa, Asia, South and Southeast Asia and Latin America<sup>4,5</sup>. We introduce a regional perspective with respect to NMDs disorders, highlighting what has been learned from epidemiological differences between LMICs as well as globally, while identifying specific needs, research priorities and the opportunities for collaboration among different LMICs (Tables 1–4).

<sup>1</sup>Centre for Neuroscience, Indian Institute of Science, Bangalore 560012, India. <sup>2</sup>Vietnam National University, Hanoi 10000, Vietnam. <sup>3</sup>Institute for Biochemical Research and School of Medicine, National University of La Plata, CC455, La Plata, 1900, Argentina. <sup>4</sup>Western Psychiatric Institute and Clinic, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania 15213, USA. <sup>5</sup>Department of Psychiatry, Mansoura University School of Medicine, Mansoura City, 35516, Egypt. <sup>6</sup>Department of Psychiatry and Human Genetics, University of Pittsburgh, Pittsburgh, Pennsylvania 15213, USA. <sup>7</sup>Department of Human Biology, Faculty of Health Sciences, University of Cape Town, Observatory 7925, South Africa. <sup>8</sup>Institute of Mental Health, Peking University, Beijing 100191, China. Correspondence should be addressed to V. R. e-mail: [viji@cns.iisc.ernet.in](mailto:viji@cns.iisc.ernet.in).



## SUB-SAHARAN AFRICA

Malnutrition, from birth through to adulthood, seems to be the most significant contributor to disease burden and disability in sub-Saharan Africa<sup>6</sup>. Maternal malnutrition, including micronutrient deficiencies such as vitamins and iodine, impairs the development and function of the nervous system of offspring, and negative effects can persist in the next generation<sup>6</sup>. Other forms of maternal and environmental trauma during the perinatal period affect brain development and cause long-term changes in brain function. Neurological disorders caused by eating toxic foodstuffs are unique to sub-Saharan Africa. Cassava is an important food crop that contains endogenous neurotoxins and, if not properly prepared, can cause konzo — a peripheral polyneuropathy with prominent sensory loss and ataxia. Lathyrism that presents as spastic paraparesis is an equally debilitating neurological disorder caused by excessive ingestion of the grass pea *Lathyrus sativus* that contains the excitotoxic amino acid,  $\beta$ -N-oxalyl amino-L-alanine<sup>6</sup>.

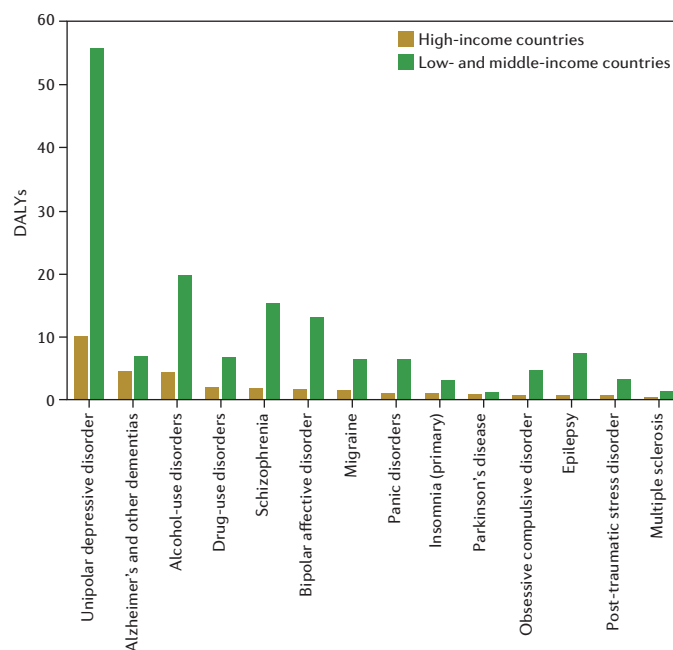
Use of psychostimulants is another major contributor to the burden of brain disorders in sub-Saharan Africa<sup>7</sup>. Of particular concern is the high prevalence of maternal alcohol and methamphetamine use in areas such as the Western Cape Province of South Africa. The incidence of fetal alcohol syndrome in some locations within this region is the highest in the world<sup>8</sup>. The increase of methamphetamine use in pregnant women in the Western Cape is of concern given the negative effects that the drug has on the developing fetus<sup>9</sup>. Khat use is of concern in East Africa<sup>10,11</sup>, where 60–90% of men use the drug daily<sup>12,13</sup>. The consequences of habitual khat consumption include behavioural disturbances and toxic psychosis, which has a particular impact on the overall health of young adults.

The prevalence and incidence of epilepsy in sub-Saharan Africa countries is twofold higher than that of other countries<sup>14–19</sup>. The prevalence varies between 4.5 and 20.8 per 1,000 people, owing to the localized and high incidence of parasitic infections, poor perinatal care and poor access to treatment. The full burden of epilepsy in sub-Saharan Africa is difficult to assess and is likely to be under-reported because people with epilepsy are stigmatized and frequently left untreated<sup>19</sup>. Stroke is another concern among non-communicable disorders within sub-Saharan Africa — incidence is increasing at an alarming rate<sup>20</sup>. The prevalence of dementia in sub-Saharan Africa is reportedly much lower than in other regions<sup>21,22</sup>. However, these reports may not be a true reflection of the prevalence, which it is projected to increase with an increase in lifespan. Furthermore, as research extends into rural areas, diagnosis of unreported cases may reveal the true burden.

Sub-Saharan Africa has the highest burden of infectious diseases and the poorest public health infrastructure in the world<sup>6,23</sup>. Parasitic infections are also highest in this region and often have neurocognitive sequelae. HIV-associated neurological disorders are a major burden, with more than 1.5-million children living with HIV and at risk of developing HIV-associated cognitive impairment and dementia<sup>1,24</sup>. Little is known of the effects of HIV and antiretroviral treatment on the developing brain. There is an urgent need for research on the longitudinal trajectory of neurodevelopment among children and adolescents who are perinatally infected with HIV<sup>24</sup>. Cognitive and psychiatric problems have been found to decrease antiretroviral treatment adherence and survival of adults with HIV in Zambia<sup>25</sup>. Neuroimaging and neurocognitive testing are well established in several regions within sub-Saharan Africa and have been used in cross-country collaborations to further our understanding of the spectrum of neurocognitive disorders in patients with HIV and to determine the effect of antiretroviral therapy on these individuals<sup>26</sup>. Subtle changes in white-matter integrity have been used for early diagnosis and monitoring progression of neurological disease in individuals with HIV<sup>26</sup>.

## MIDDLE EAST AND NORTH AFRICA

Many of the aetiological and treatment features of psychiatric disorders in the Middle East and North Africa are due, in part, to the unique environmental and cultural influences within the region. Over the past few



**Figure 1** | Comparison of disability associated life years (DALYs) between high-income and low- and middle-income countries. The data were derived from the World Health Organization ([http://www.who.int/healthinfo/global\\_burden\\_disease/metrics\\_daly/](http://www.who.int/healthinfo/global_burden_disease/metrics_daly/)) and refs 4, 5.

decades, communities have been exposed to traumatic events including anti-government uprisings and wars, which has left many populations vulnerable to mood disorders, such as post-traumatic stress disorder (PTSD) and major depressive disorder (MDD). In comparison with the global estimate of 4.4% (ref. 27), depression prevalence in Iraq is 7.2% and is 15.3% in the Palestinian territories<sup>28,29</sup>. In fact, MDD is currently listed among the top three causes of YLDs in most of the countries within the Middle East and North Africa<sup>2</sup>. The statistics are similar for PTSD within the region.

Owing to the high rate of consanguinity in the region, the incidence of several recessively inherited genetic disorders, such as inherited deafness, is increasing<sup>30–32</sup>. For example, Bardet-Biedl syndrome, which includes many nervous system abnormalities, is common in most of the Arab countries, particularly in Kuwait. Whereas the syndrome typically affects 1 in 150,000 people in North America and Europe, the prevalence in Arab countries ranges from 1 in 13,500 to 1 in 30,000 people<sup>30</sup>. A national strategy is needed in this region to address this burden of genetic disease. Although services such as genetic screening exist, understanding the barriers to access and use requires implementation research and an understanding of sociocultural norms. This will help health workers to tailor services and educational campaigns that are culturally acceptable.

The prevalence of substance-use disorders varies between 7.25% and 14.5%, with cannabis being the most commonly used drug followed by alcohol<sup>12,33</sup>. Khat is also widely used as a stimulant in Yemen and the neighbouring countries within the Arabian Peninsula.

There is a need for population-based prevalence estimates of common neurological disorders in the Middle East and North Africa, with a special emphasis on epilepsy, because systematic epidemiological studies of epilepsy in Asia and Africa have not included this region<sup>34</sup>. Most published studies only report hospital-based samples<sup>35</sup>. For example, a review of seizure disorders in Arab countries indicated a median prevalence of 2.3 per 1,000 people (range, 0.9–6.5 per 1,000). These figures are very likely to underestimate the prevalence in a population of more than 350-million people<sup>36</sup>, particularly because epilepsy is stigmatized within several communities<sup>37</sup>.

## LATIN AMERICA AND THE CARIBBEAN

Within the countries and territories of Latin America and the Caribbean (Central America, Mexico and the Latin Caribbean); the non-Latin Caribbean and South America there are sub-regional differences in the contribution of NMDs disorders to the total burden of disease measured in DALYs. Although DALYs owing to neurological disorders, including stroke, are low in the Andean Latin American sub-region, they are higher in the southern Caribbean sub-regions and even higher in tropical Latin America and the Caribbean. However, if one considers the total region, the burden of NMDs disorders accounts for 22.2% of the total DALYs. The overall weighted prevalence of mental health disorders in children in the region (12.7%) is significantly more than the prevalence (9.7%) seen in United Kingdom when similar diagnostic procedures are used<sup>38</sup>. Importantly, there is inadequate information on risk and protective factors that affect the incidence of mental health disorders in children living in developing countries in general and Latin America and the Caribbean in particular<sup>39</sup>.

Unipolar depressive disorders (13.2%) and alcohol dependence (6.9%) constitute the most common psychiatric disorders<sup>40</sup> in Latin America and the Caribbean (Fig.1). The annual level of alcohol consumption (8.4 litres per capita annually) is the second highest in the world after Europe<sup>41</sup>. Alcohol consumption has been associated with roughly a third of intentional and non-intentional accidents<sup>42</sup>; traumatic brain injuries incurred from any type of accident have long-term implications for society and for the individual, including impaired attention, depression and economic costs to families<sup>43</sup>.

As for other regions the current increasing trend in DALYs for non-communicable disorders<sup>2</sup> suggests that epilepsy and dementia are unique in terms of their increasing prevalence. Their prevalence or manifestation is increasing in Latin America and the Caribbean. The annual incidence of epilepsy according to a collection of 32 community-based studies is 77.7 to 190 per 100,000 people each year<sup>44</sup>, compared with 30 to 50 per 100,000 people in high-income countries. Distribution of epilepsy across sub-regions of Latin America and the Caribbean also differs; one reason for this is the direct association between epilepsy and the distribution of neurocysticercosis<sup>45</sup>. Dementia is also widespread<sup>46,47</sup>, but pockets of early onset Alzheimer's disease in families are apparent in Caribbean Hispanic people who originate from Puerto Rico or the Dominican Republic<sup>21</sup>. Studies on familial types of dementia in Latin American countries such as Colombia (Alzheimer's disease) and Venezuela (Huntington's disease) have shown that both non-genetic (nurture) and unrelated genetic factors may have a major role in influencing phenotypes<sup>48–50</sup>. This suggests that even highly penetrant autosomal dominant diseases may be modified by environment or lifestyle factors. Although not unique to the region, it is worth noting that stroke is the leading cause of death in Ecuador, and in other Latin American countries<sup>51</sup>. Little is known about the prevalence of any of these disorders among indigenous Andean or Amazonian populations.

## ASIA

Sub-regions of Asia comprise East and Southeast Asia, and incorporate the Association of Southeast Asian Nations as well as China, whereas South Asia consists of sub-Himalayan countries, including Afghanistan, Bangladesh, India, Pakistan and Sri Lanka. About two-thirds of the world's population resides on the Asian continent. India and China, because of their size and economic impact, have a major influence on the health and trends of the region, and in shaping global health statistics, however they are catalogued. Asia's ethnic diversity, and widely disparate socioeconomic development lead to significant variations in the prevalence and burden of NMDs disorders. An epidemiological study<sup>52</sup> of epilepsy in 23 Asian countries revealed the lifetime prevalence of epilepsy to be 1.5 to 14 per 1,000 people. Infections of the nervous system often contribute to epilepsy and prevention of these infections is needed to reduce the burden of the condition.

Another major concern is the rising prevalence of dementia; although the number of patients with dementia is predicted to increase

by 100% between 2001 and 2040 in developed countries, dementia is predicted to increase by more than 300% in India, China, South Asia and the Western Pacific region<sup>21</sup>. In India alone, there are 3.7 million people with dementia and the numbers are expected to double by 2030 (ref. 53). In addition, the high burden of cardiovascular risk factors in developing countries, including India, contributes to cerebrovascular disease such as vascular dementia<sup>54</sup>.

Asia, in particular South Asia, has the highest stroke mortality in the world<sup>55</sup>. Within Asia, there is a wide variation in stroke prevalence<sup>56</sup>. Rural parts of South Asia have lower stroke prevalence than urban areas<sup>56</sup>, and this needs to be examined further in future research<sup>57</sup>. In China, the incidence of stroke differs geographically. A higher incidence of stroke is seen in northern and western areas, and is associated with a higher prevalence of hypertension and obesity<sup>58</sup>. Barriers to preventing and reducing mortality and disability due to stroke are the lack of infrastructure, such as dedicated stroke care units, and awareness<sup>57</sup>.

Tobacco use — a leading cause of stroke — is a major public health issue for East and Southeast Asia. Half of the world's tobacco consumption takes place in Asia<sup>59</sup>. Men are more likely to smoke than women; and prevalence rates for males range across countries from 36% in Singapore to 64% in Laos<sup>60</sup>. Although the neurological and other health implications of smoking are well known, many Asian people still smoke. Public health measures to reduce smoking are just beginning; for example, in June 2005 and October 2008, India and Beijing banned indoor smoking in public places and offices, respectively.

## COMMON RESEARCH NEEDS AND CHALLENGES

There are several commonalities within LMICs in terms of disease prevalence and the public health and research challenges, although considerable ethnic and geographical diversity exists.

### Lack of robust epidemiological studies

Epidemiological studies, preferably longitudinal, designed to identify disease burden and risk or protective factors for NMDs disorders, are one of the most important research needs in LMICs. These need to be complemented by research on health systems and sociocultural effects, and clinical trials to determine the best interventional strategies. Furthermore, rapid urbanization and the associated demographic and sociocultural changes in LMICs should be studied with respect to their impact on the course and outcome of different brain disorders, especially mental health illnesses and substance misuse. A careful analysis of the possible interaction between demographic and sociocultural changes, and biological factors is essential to initiate remedial steps to contain the progression of these disorders.

### Disproportionate distribution of scientists

Some countries have a disproportionate share of scientists, with investment and output concentrated in only a few places. In general, Latin America produces more neuroscience and mental health disorder publications than the Middle East and Africa. Similar variation is seen in the number of neuroscience publications produced in Asia (Fig. 2). Between 1996 and 2013, India consistently produced the most neuroscience and mental health research publications. Figures also reveal that 9.2% of institutions in India produce 80.1% of the publications. Among Latin American and Caribbean countries, Brazil now accounts for more than two-thirds of South America's entire research output, although in terms of articles per capita, it is broadly similar to Argentina, Uruguay and Chile. One could leverage this situation by promoting intraregional research collaborations to enhance research capacity and infrastructure. The top 10 African countries in terms of health-research publications from 2000 to 2014 are South Africa, Nigeria, Kenya, Uganda, Tanzania, Ethiopia, Ghana, Cameroon, Malawi and Senegal<sup>61</sup>. Although these trends comprised all health research, it is likely that mental health publications are ranked similarly in sub-Saharan Africa.

**Table 1** | Neurological, mental health, developmental and substance-use disorders and specific burden of disease in sub-Saharan Africa

Condition or disease	Key affected countries	Burden of disease	Impact of condition or disease
<b>Nutrition: malnutrition</b>	All SSA	<ul style="list-style-type: none"> <li>204 million people suffer from hunger<sup>80-82</sup></li> <li>Highest prevalence of stunting in the world is in East Africa (42%) and West Africa 36% based on the WHO Child Growth Standards<sup>83</sup></li> <li>22% of children are underweight in West Africa<sup>83</sup></li> </ul>	<ul style="list-style-type: none"> <li>Maternal malnutrition impairs the development and function of the nervous system of offspring and negative effects persist in the next generation<sup>6</sup></li> <li>Malnutrition in infants and children affects their growth and cognitive development<sup>6,84</sup></li> </ul>
<b>Nutrition: the toxic nutritional neurological disorders konzo (cassava) and lathyrism (grass pea)</b>	Cameroon, Central African Republic, Democratic Republic of Congo, Tanzania, Ethiopia and Mozambique  Ethiopia	<ul style="list-style-type: none"> <li>Reported estimates show there are around 6,500 cases of cassava toxicity; unofficial reports estimate the number of cases to be at least 100,000 (ref. 85)</li> </ul>	<ul style="list-style-type: none"> <li>Leads to difficulty in walking<sup>84</sup> and peripheral polyneuropathy with prominent sensory loss and ataxia<sup>6</sup></li> <li>Malnutrition may increase the negative impact of food borne toxins and causes irreversible spasticity<sup>86</sup></li> </ul>
<b>Substance use: cannabis, methamphetamine, khat, alcohol, and opioids or heroin</b>	West and Central Africa (notably), and South Africa  South Africa  Tanzania, Kenya, Uganda, Ethiopia, Eritrea and Somalia  South Africa  West and Central Africa and South Africa	<ul style="list-style-type: none"> <li>Cannabis use is higher than the global average (12.4% versus 3.8%)<sup>87</sup></li> <li>Cannabis is the most popular illicit drug followed by cocaine</li> <li>Increased use of methamphetamine during pregnancy<sup>8,9,88</sup></li> <li>General increase in drug use<sup>87</sup></li> <li>60–90% of East African males use khat daily<sup>10-13,89</sup></li> <li>Self-reported prevalence of alcohol abuse is 36.9%</li> <li>Fetal alcohol syndrome in the local Western Cape population is the highest in the world<sup>8</sup></li> <li>Annual prevalence of heroin use is above the global average<sup>88</sup></li> <li>0.92–2.29 million people used opiates in the past year<sup>87</sup></li> </ul>	<ul style="list-style-type: none"> <li>1 in 18 problem drug users receive treatment; most of those in treatment are cannabis users<sup>87</sup></li> <li>Structural (volume reductions in the striatum and increases in limbic areas of the brain) and functional deficits as well as cognitive and behavioural abnormalities have been described in infants and children exposed to methamphetamine prenatally<sup>9</sup></li> <li>Violent behaviour in adults</li> <li>Cognitive dysfunction<sup>89</sup></li> <li>Chronic khat use may have a long-term deleterious effect on working memory<sup>80</sup></li> <li>Negative effects on the developing fetus<sup>9</sup></li> <li>Fetal alcohol syndrome, growth retardation and cognitive dysfunction<sup>8</sup></li> <li>An increasing role as a transit area for drug trafficking and increased crime rate<sup>87</sup></li> </ul>
<b>Epilepsy</b>	All SSA	<ul style="list-style-type: none"> <li>Prevalence varies between 4.5 and 20.8 per 1,000 people; about twice that elsewhere<sup>14-19</sup></li> </ul>	<ul style="list-style-type: none"> <li>Impaired cognitive function due to effect of seizures on the developing brain<sup>91</sup></li> <li>Stigma and social isolation<sup>19</sup></li> </ul>
<b>Stroke</b>	All SSA	<ul style="list-style-type: none"> <li>Community-based studies revealed an age-standardized annual stroke incidence rate of up to 316 per 100,000 of the population, and age-standardized prevalence rates of up to 981 per 100,000 (ref. 92)</li> <li>65% of all neurological admissions to hospitals are stroke related in the West African sub-region<sup>92</sup></li> </ul>	<ul style="list-style-type: none"> <li>Increased burden to society</li> </ul>
<b>Dementia</b>	Nigeria, Democratic Republic of Congo, Senegal, Central African Republic, Tanzania, Zambia and Kenya	<ul style="list-style-type: none"> <li>Prevalence is between &lt;1% and 10.1% in population-based studies and up to 47.8% in hospital-based studies<sup>93</sup></li> </ul>	<ul style="list-style-type: none"> <li>A burden to family and society</li> </ul>
<b>HIV-associated neurological conditions</b>	South Africa, Kenya, Nigeria, Zambia, Malawi, Cameroon, Botswana and Uganda,	<ul style="list-style-type: none"> <li>1.5-million children are living with HIV and are at risk of developing HIV-associated cognitive impairment and dementia<sup>3,24</sup></li> <li>Prevalence of HIV-related neurocognitive impairment ranged from &lt;1% to 80% in hospital-based studies<sup>93</sup></li> </ul>	<ul style="list-style-type: none"> <li>HIV-related dementia is a particular concern, and burden, in SSA as people live longer with the disease</li> <li>Children infected with HIV perinatally do not perform as well as non-infected children on cognitive tests and are at much higher risk for psychiatric disorders later in life<sup>24</sup></li> </ul>

SSA, sub-Saharan Africa; WHO, World Health Organization. The 2014 population estimates for sub-Saharan Africa were 961.5 million (<http://data.worldbank.org/region/SSA>)

## Insufficient resources for treatment and research

Most countries allocate less than 5% of their health-care budget to the treatment of brain disorders<sup>62,63</sup>. For example the Middle East and North Africa, Palestine, Qatar and Egypt, spend only 2.5%, 1% and less than 1% on brain-disorder treatment, respectively<sup>64</sup>. The number of mental health professionals available in most LMICs is also very low. For example, there are only 1.44 psychiatrists per 100,000 people in Egypt. In India, 52% of the districts do not have psychiatric facilities, and there is an acute shortage of psychiatrists, psychologists and psychiatric social workers<sup>65</sup>. Hence people with neuropsychiatric disorders remain largely undiagnosed and even when they are diagnosed, they do not have access to sustainable, affordable treatment and optimal medical care<sup>66</sup>. Although a recent World Bank report indicates that disease burden that results from non-communicable causes, including mental health disorders, has increased substantially, with major depressive disorders at the top of the list (<http://www.healthdata.org/gbd/data>)<sup>2</sup>

there is a severe lack of resources, particularly of trained personnel and training facilities<sup>67</sup>. Given the severe fiscal and human-resource constraints for treatment, it is not surprising that research is lagging. The current research gap between developed and developing nations is reflected in the mental health research output, with LMICs contributing to only 6% of international research articles<sup>68</sup>.

## Brain drain

Brain drain is the loss of highly trained people, constituting another big challenge to LMICs, and widening the research gap between high-income countries and LMICs. The reasons cited by researchers for their exodus are a dearth of funding, poor facilities, and limited or a lack of peer groups to provide intellectual stimulation<sup>69</sup>. Although it may be argued that brain drain is a common problem in LMICs across disciplines, neuroscience research is particularly affected. This is because unlike core disciplines such as chemistry, physics or mathematics,



**Table 2 |** Neurological, mental health, developmental and substance-use disorders and specific burden of disease in South Asia and Southeast Asia

Condition or disease	Key affected countries	Burden of disease	Impact of condition or disease
<b>Mood disorders</b>	Vietnam, Cambodia and South Asia countries	<ul style="list-style-type: none"> <li>Depression is the second most common NMDs disorder in Vietnam (2.8% prevalence) and Cambodia (16.7% prevalence); there is a relatively high prevalence (23.6%) in elderly Chinese<sup>69,94,95</sup></li> <li>Anxiety is the most common NMDs disorder in Cambodia (27.4% prevalence)<sup>2,5</sup></li> <li>Prevalence of 16/1,000 population in India<sup>96</sup></li> <li>Unipolar depression ranks among the top 10 disorders<sup>2</sup></li> </ul>	<ul style="list-style-type: none"> <li>Substantial impact on society in general and family in particular</li> <li>Patients with psychiatric disorders under diagnosed and undertreated due to scarcity of physicians<sup>97</sup> coupled with absence of evidence for effectiveness of treatment<sup>98</sup></li> <li>People with dysthymia have impaired quality of life and poor marital adjustment<sup>99</sup></li> </ul>
<b>Dementia</b>	All Asia	<ul style="list-style-type: none"> <li>Predicted to increase by more than 300% in India, China, South Asia and the Western Pacific region<sup>11</sup></li> <li>9 million Chinese have dementia<sup>100</sup></li> <li>The rate in people over 60 was 3.4% in Thailand and 3.5% in Indonesia<sup>101</sup></li> <li>China and India are predicted to have the largest number of dementia cases in the next decade<sup>21,100</sup></li> <li>An estimated 3.7 million Indians have dementia and the numbers are expected to double by 2030 (ref.102)</li> </ul>	<ul style="list-style-type: none"> <li>People with dementia who live with families puts significant burden on carers. None of the carers receive carer benefits and have high levels of psychological morbidity<sup>102</sup></li> <li>Annual cost of dementia in China is US\$2,384 per patient annually<sup>103</sup></li> </ul>
<b>Stroke</b>	All Asia	<ul style="list-style-type: none"> <li>Prevalence of 45–471 per 100,000 people in South Asia<sup>6</sup></li> <li>Annual stroke mortality in China is 1.6 million, approximately 157 per 100,000, which has exceeded heart disease as the leading cause of death and adult disability<sup>104</sup></li> <li>Among a sample of five ASEAN countries (Indonesia, Myanmar, Vietnam, Thailand and Malaysia), stroke was the top cause of death<sup>105</sup></li> <li>Stroke mortality in South Asia is the highest in the world, accounting for more than 40% of global stroke deaths<sup>97</sup></li> <li>Mortality in South Asia is 73 per 100,000 of the population<sup>97</sup></li> </ul>	<ul style="list-style-type: none"> <li>Leading cause of death, long-term disability</li> <li>Incidence of stroke differs geographically in China — there is a higher incidence in northern and western areas, which are associated with higher prevalence of hypertension and obesity<sup>58</sup></li> <li>Rural parts of South Asia have a lower stroke prevalence compared with urban areas<sup>56</sup></li> <li>There are less than 100 stroke care units in South Asia<sup>97</sup>, leading to poor care for patients and increased morbidity and mortality</li> <li>Barriers to stroke thrombolysis in South Asia include a lack of infrastructure, lack of awareness and a lack of affordability<sup>97</sup>, leading to increase in morbidity and mortality</li> </ul>
<b>Traumatic brain injury</b>	All Asia	<ul style="list-style-type: none"> <li>44% of the world's road deaths occur in Asia<sup>106</sup></li> <li>The incidence rate of TBI in India is 160 per 100,000; 1.6 million people will sustain a TBI<sup>106</sup></li> </ul>	<ul style="list-style-type: none"> <li>India has the highest rate of TBI due to falls, and accounts for 50% of global falls<sup>106</sup></li> </ul>
<b>Tobacco use</b>	East and Southeast Asia	<ul style="list-style-type: none"> <li>Prevalence of male smokers ranges from 36% in Singapore to 64% in Laos<sup>60</sup>; the rate is lower in women<sup>58</sup></li> <li>Smoking in children aged between 13 and 15 is common in ASEAN<sup>60</sup></li> <li>Half of the world's tobacco is consumed in Asia<sup>59</sup></li> <li>South Asia has the highest use of smokeless tobacco worldwide<sup>59</sup></li> </ul>	<ul style="list-style-type: none"> <li>Chronic nicotine consumption induces neuro-adaptations in the brain's reward system that result in nicotine dependence<sup>107</sup></li> <li>Withdrawal from nicotine can include somatic symptoms (for example, jumping, shaking, abdominal constrictions, chewing, scratching, and facial tremors) or affective symptoms (for example anhedonia)<sup>107</sup></li> <li>Past smokers are prone to relapse for weeks, months or even years after cessation<sup>107</sup></li> <li>Nicotine affects mood and cognition by stimulating nicotinic acetylcholine receptors on neurons in the brain's mesolimbic reward system<sup>107</sup></li> </ul>

ASEAN, Association of Southeast Asian Nations; NMDs, neurological, mental health, developmental and substance-use; TBI, traumatic brain injury. The 2014 population estimates for South Asia were 1.692 billion and those for East Asia and Pacific region were 2.02 billion

neuroscience is an interdisciplinary field and most LMICs do not have adequate training capacity. This, combined with the fact that the expensive infrastructure needed for some areas of brain research is often not available, drives many researchers from LMICs to migrate to high-income countries.

## REGION SPECIFIC RESEARCH NEEDS AND CHALLENGES

There are specific needs across the regions that constitute LMICs, which have to be addressed in a region- and/or country specific manner.

### Identification of risk and protective factors

There is an immediate need to characterize population groups that have increased susceptibility or resilience to brain disorders or better clinical outcomes, which could lead to the identification of disease-modifying factors and interventions in other populations. Opportunities for research have been observed in different regions. For example, the course and outcome of schizophrenia is better understood in India than in other countries<sup>70</sup>. The lifetime prevalence of PTSD as a major depressive disorder is not significantly greater in Southeast Asia compared with other parts of the world<sup>1</sup>, despite the region being a natural disasters-prone region. As a region with significant population growth trends, the likely increase in the number of people with childhood and adolescent disorders (including learning disabilities) at one end of the spectrum and increasing lifespan

that leads to higher incidence of age-related neurodegenerative disorders (including dementia) at the other end make it imperative that resources are channelled to research aimed at identifying risk and protective factors<sup>5,71–73</sup>.

### Integration of traditional methods of treatment

Assessing the efficacy of indigenous, traditional Chinese medicine and Indian Ayurveda medicine for brain disorders is important. Integrating traditional Buddhist practices in the treatment of psychiatric disorders, such as the integration of mindfulness techniques into cognitive behavioural therapy, has created new intervention approaches including mindfulness-based cognitive therapy<sup>74</sup>, mindfulness-based stress reduction<sup>75</sup>, and dialectical behaviour therapy<sup>76</sup>. Similarly yoga, as an addition to pharmacological interventions, is beneficial in the treatment of schizophrenia and depression<sup>77,78</sup>.

### Collaborations and knowledge generation

Opportunities have been made possible by improvements in infrastructure in sub-Saharan Africa, which sets the stage for cross-country collaboration. For example, in addition to South Africa, several countries have neuroimaging facilities, which can be used to analyse brain structure and function to aid diagnosis and treatment. Malawi has excellent electroencephalography (EEG) services and the capacity to conduct longitudinal studies. Zambia has very good imaging

**Table 3 |** Neurological, mental health, developmental and substance-use disorders and specific burden of disease in Latin America and the Caribbean

Condition or disease	Key affected countries	Burden of disease	Manifestation of condition or disease
Unipolar depressive disorder	All LAC	<ul style="list-style-type: none"> <li>One of the most common mental illness<sup>39</sup>, constituting 13.2% of the burden of total DALYs</li> <li>Represents 35.7% among psychiatric disorders and is more prevalent among lower-income groups<sup>105</sup></li> </ul>	<ul style="list-style-type: none"> <li>Typical manifestations are irritability, difficulty with concentration, fatigue or lack of energy, feelings of hopelessness and/or helplessness and sleep problems</li> </ul>
Substance use: alcohol and tobacco	All LAC  All LAC (higher in South America)	<ul style="list-style-type: none"> <li>One of the most common psychiatric disorders<sup>39</sup>, constituting 6.9% burden of total DALYs</li> <li>Chile has the highest consumption rate of alcohol and tobacco<sup>40</sup></li> <li>Prevalence of smoking in men is 31% and in women is 17%<sup>109</sup></li> </ul>	<ul style="list-style-type: none"> <li>Alcohol consumption is a trigger of violence and accidents and is associated with 33% of intentional accidents and 26% of non-intentional accidents<sup>41</sup></li> <li>Both associated with acute as well as long-term chronic conditions that range from brain damage, high blood pressure and stroke to liver and muscle diseases<sup>108</sup></li> </ul>
Traumatic brain injury	All LAC	<ul style="list-style-type: none"> <li>The region has one of the highest rates of injury mortality in the world<sup>110–112</sup></li> <li>Has the highest incidence rates of traumatic brain injury caused by violence<sup>106</sup></li> </ul>	
Epilepsy	Honduras, Panamá, Chile, Peru and Colombia	<ul style="list-style-type: none"> <li>17.8 (range, 6–43.2) per 1,000 people</li> <li>Incidence is 77.7–190 per 100,000 people per year<sup>44</sup>, whereas in high-income countries it is 30–50 per 100,000</li> </ul>	<ul style="list-style-type: none"> <li>Epilepsy is characterized by the appearance of primary generalized or partial seizures that begin with a widespread electrical discharge that involves one or both sides of the brain at once. Hereditary factors are important in many of these seizures</li> </ul>
Dementia	All LAC including large family groups in the Dominican Republic, Colombia, and Venezuela	<ul style="list-style-type: none"> <li>6% of those over 60 years are affected<sup>21</sup></li> <li>Affects 2 million people, and likely to increase<sup>47</sup></li> </ul>	

DALYs, disability adjusted life years; LAC, Latin America and the Caribbean. The 2014 population estimates for Latin America were 521.9 million and for the Caribbean were 7.0 million

**Table 4 |** Neurological, mental health, developmental and substance-use disorders and specific burden of disease in the Middle East and North Africa

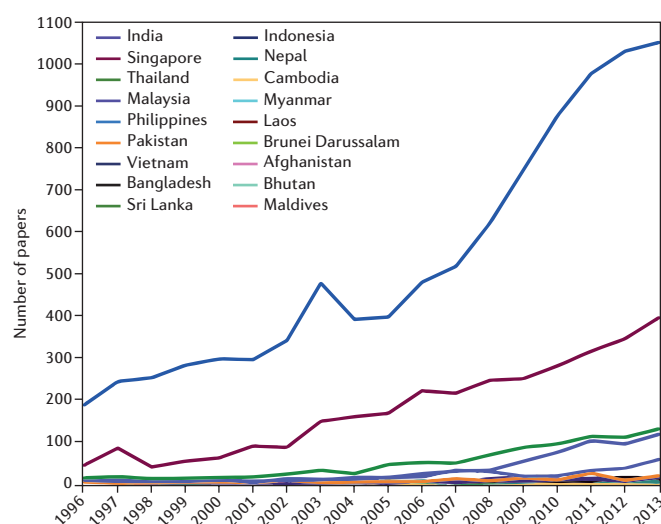
Condition or disease	Key affected countries	Burden of disease	Impact of condition or disease
Unipolar depressive disorder	All countries	<ul style="list-style-type: none"> <li>Prevalence between 5–10%<sup>113</sup></li> <li>15.3% Palestinian adults and children have depression<sup>28</sup></li> <li>The lifetime prevalence in Iraq is 7.2%<sup>27</sup></li> </ul>	<ul style="list-style-type: none"> <li>Women are more likely to have higher rates of depression than men</li> <li>Disability, marital dysfunction, loss of employment and risk of suicide<sup>113</sup></li> </ul>
Post-traumatic stress disorder	Conflict zone countries	<ul style="list-style-type: none"> <li>36% of adults in Iraq suffer from psychological trauma as a result of violence<sup>114</sup></li> <li>A rate of 23.2% has been reported in Palestinian populations in the Gaza strip and Nablus district in the West Bank<sup>28</sup></li> <li>Prevalence of 37.1% for Palestinian children<sup>115</sup></li> </ul>	<ul style="list-style-type: none"> <li>Disability, loss of employment, disrupted family relationships and risk of substance misuse</li> </ul>
Substance-use disorders (including nicotine, cannabis, alcohol and opiates)	All countries	<ul style="list-style-type: none"> <li>80% of men and 67.8% of women in Yemen have used khat during their lifetime</li> <li>There has been an increase in the prevalence of substance use in the Arabian Peninsula and East Africa, particularly among young adults and females</li> <li>Tramadol use is a serious, growing public health problem in Egypt and other Middle East and North African countries (8.8% use among school children in Egypt)<sup>116</sup></li> <li>Cannabis is the most commonly used drug</li> <li>In a 10 country study the tobacco smoking rate was 31.2%. The highest rates were in Jordan, Lebanon, Syria and Turkey<sup>117</sup></li> </ul>	<ul style="list-style-type: none"> <li>Khat is implicated in depression, anxiety, psychosis and cognitive dysfunction<sup>91,118</sup></li> <li>Early first drug use leads to more drug problems later in life<sup>119</sup></li> </ul>
Recessively inherited genetic diseases	All countries	<ul style="list-style-type: none"> <li>Incidence is related to high consanguinity rates<sup>30–32</sup></li> </ul>	<ul style="list-style-type: none"> <li>Increased need for medical care</li> <li>Reduced lifespan</li> <li>Increased family burden<sup>30,31</sup></li> </ul>
Epilepsy	All countries	<ul style="list-style-type: none"> <li>Median prevalence is estimated to be 2.3 per 1,000 (ref. 36)</li> <li>In 23 Asian countries lifetime prevalence of epilepsy was 1.5 to 14 per 1,000 (ref. 54)</li> </ul>	Prevalence is likely to be underreported because of stigma associated with the illness

The 2014 population estimates Middle East and North Africa were 351.4 million. Population data from World Bank and aggregated by <http://data.okfn.org/data/core/population>

and neurophysiology (EEG and nerve conduction velocity) facilities for adults and children, as well as the capacity for population-based studies in rural and urban centres and longitudinal cohort studies. In South Africa, a wide range of research techniques have been developed, including EEG, electromyography, magnetic resonance imaging, diffusion tensor imaging, structural imaging, magnetic resonance spectroscopy, positron emission tomography and transcranial magnetic stimulation.

## Health budgets and research funding

A lack of adequate funding opportunities for neuroscience research in LMICs is a major hindrance to moving the field forward. The disproportionate designation of health spending in relation to variable national gross domestic product in LMICs makes it difficult to sustain or even designate research budgets<sup>23</sup>. For example, the order of the top three countries in sub-Saharan Africa — South Africa, Nigeria and Kenya — in terms of health research publications has remained unchanged for



**Figure 2** | Number of neuroscience papers in international peer-reviewed journals published by authors from Asian countries per year. The data were retrieved from <http://www.scimagojr.com>.

the past 14 years, because of financial constraints imposed by total expenditure on health and the national gross domestic product<sup>61</sup>. Funding for NMDs disorders research is variable and depends on the priorities of the government agencies that fund health and/or science and technology research in general (where these exist). Three steps could be taken to promote neuroscience research in LMICs. First, governmental funding for research through universities and research institutions should be enhanced and encouraged. Second, funds from national and international non-governmental organizations (NGOs; which contribute up to 20% of all external aid for health services in developing countries, <http://www.imva.org/Pages/biblfrm.htm>) could be used to increase research opportunities in health and medicine, including epidemiology, clinical research, public health services and policy research. Third, increased collaboration with regional or international partners could lead to more research opportunities and support.

## CONCLUSIONS

Regional variations in the challenges posed by NMDs disorders among LMICs means that research priorities need to be addressed country-by-country, and by regions within countries. There are significant gaps between the resources needed for research and those that are currently available, and a pressing need to strengthen human-resource capacity and research infrastructure, while promoting collaboration. Global demographic trends point to LMICs as the main work force of the future<sup>79</sup>; it is, therefore, imperative to act expeditiously to reduce the enormous burden of brain disorders in these countries. The loss of human potential and cost of inaction are unacceptably high.

1. Murray, C. J. et al. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* **380**, 2197–2223 (2012).
2. Global Burden of Disease Study Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* **386**, 743–800 (2015).
3. Murray, C. J. & Lopez, A. D. Regional patterns of disability-free life expectancy and disability-adjusted life expectancy: global Burden of Disease Study. *Lancet* **349**, 1347–1352 (1997).
4. Whiteford, H. A. et al. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *Lancet* **382**, 1575–1586 (2013).
5. Whiteford, H. A., Ferrari, A. J., Degenhardt, L., Feigin, V. & Vos, T. The global burden of mental, neurological and substance use disorders: an analysis from the Global Burden of Disease Study 2010. *PLoS ONE* **10**, e0116820 (2015).
6. Kerac, M. et al. The interaction of malnutrition and neurologic disability in Africa. *Semin. Pediatr. Neurol.* **21**, 42–49 (2014).

7. Degenhardt, L., Whiteford, H. & Hall, W. D. The Global Burden of Disease projects: what have we learned about illicit drug use and dependence and their contribution to the global burden of disease? *Drug Alcohol Rev.* **33**, 4–12 (2014).
8. Hess, A. T. et al. A comparison of spectral quality in magnetic resonance spectroscopy data acquired with and without a novel EPI-navigated PRESS sequence in school-aged children with fetal alcohol spectrum disorders. *Metab. Brain Dis.* **29**, 323–332 (2014).
9. Kwiatkowski, M. A., Roos, A., Stein, D. J., Thomas, K. G. & Donald, K. Effects of prenatal methamphetamine exposure: a review of cognitive and neuroimaging studies. *Metab. Brain Dis.* **29**, 245–254 (2014).
10. Odenwald, M. in *Neglected Tropical Disease and Conditions of the Nervous System* (eds Bentivoglio, M. et al.) 293–306 (Springer, 2014).
11. Patel, N. B. in *Neglected Tropical Disease and Conditions of the Nervous System* (eds Bentivoglio, M. et al.) 307–320 (Springer, 2014).
12. Njuguna, J., Olivia, S., Muruka, C. & Owek, C. Khat consumption in Masalani town, northeastern Kenya. *J. Psychoactive Drugs* **45**, 355–359 (2013).
13. Warfa, N. et al. Khat use and mental illness: a critical review. *Soc. Sci. Med.* **65**, 309–318 (2007).
14. Ba-Diop, A. et al. Epidemiology, causes, and treatment of epilepsy in sub-Saharan Africa. *Lancet Neurol.* **13**, 1029–1044 (2014).
15. Mustapha, A. F., Preux, P. M., Sanya, E. O. & Akinleye, C. A. The prevalence and subjective handicap of epilepsy in Ilie—a rural riverine community in South West Nigeria: a door-to-door survey. *Epilepsy Behav.* **37**, 258–264 (2014).
16. Ngugi, A. K. et al. Prevalence of active convulsive epilepsy in sub-Saharan Africa and associated risk factors: cross-sectional and case-control studies. *Lancet Neurol.* **12**, 253–263 (2013).
17. Osakwe, C., Otte, W. M. & Alo, C. Epilepsy prevalence, potential causes and social beliefs in Ebonyi State and Benue State, Nigeria. *Epilepsy Res.* **108**, 316–326 (2014).
18. Wagner, R. G. et al. Prevalence and risk factors for active convulsive epilepsy in rural northeast South Africa. *Epilepsy Res.* **108**, 782–791 (2014).
19. Wilmshurst, J. M., Birbeck, G. L. & Newton, C. R. Epilepsy is ubiquitous, but more devastating in the poorer regions of the world... or is it? *Epilepsia* **55**, 1322–1325 (2014).
20. Lekoubou, A., Nkoke, C., Dzudie, A. & Kengne, A. P. Stroke admission and case-fatality in an urban medical unit in sub-Saharan Africa: a fourteen year trend study from 1999 to 2012. *J. Neurol. Sci.* **350**, 24–32 (2015).
21. Kalaria, R. N. et al. Alzheimer's disease and vascular dementia in developing countries: prevalence, management, and risk factors. *Lancet Neurol.* **7**, 812–826 (2008).
22. Mavrodaris, A., Powell, J. & Thorogood, M. Prevalences of dementia and cognitive impairment among older people in sub-Saharan Africa: a systematic review. *Bull. World Health Organ.* **91**, 773–783 (2013).
23. Sepulveda, J. & Murray, C. The state of global health in 2014. *Science* **345**, 1275–1278 (2014).
24. Laughton, B., Cornell, M., Boivin, M. & Van Rie, A. Neurodevelopment in perinatally HIV-infected children: a concern for adolescence. *J. Int. AIDS Soc.* **16**, 18603 (2013).
25. Birbeck, G. L. et al. Neuropsychiatric and socioeconomic status impact antiretroviral adherence and mortality in rural Zambia. *Am. J. Trop. Med. Hyg.* **85**, 782–789 (2011).
26. Hoare, J. et al. Systematic review of neuroimaging studies in vertically transmitted HIV positive children and adolescents. *Metab. Brain Dis.* **29**, 221–229 (2014).
27. Ferrari, A. J. et al. Burden of depressive disorders by country, sex, age, and year: findings from the global burden of disease study 2010. *PLoS Med* **10**, e1001547 (2013).
28. Alhasnawi, S. et al. The prevalence and correlates of DSM-IV disorders in the Iraq Mental Health Survey (IMHS). *World Psychiatry* **8**, 97–109 (2009).
29. Espie, E. et al. Trauma-related psychological disorders among Palestinian children and adults in Gaza and West Bank, 2005–2008. *Int. J. Ment. Health Syst.* **3**, 21 (2009).
30. Al-Gazali, L. & Hamamy, H. Consanguinity and dysmorphology in Arabs. *Hum. Hered.* **77**, 93–107 (2014).
31. Al-Gazali, L., Hamamy, H. & Al-Arriyad, S. Genetic disorders in the Arab world. *Br. Med. J.* **333**, 831–834 (2006).
32. Tadmouri, G. O., Al Ali, M. T., Al-Haj Ali, S. & Al Khaja, N. CTGA: the database for genetic disorders in Arab populations. *Nucleic Acids Res.* **34**, D602–D606 (2006).
33. Hamdi, E. et al. Lifetime prevalence of alcohol and substance use in Egypt: a community survey. *Subst. Abuse* **34**, 97–104 (2013).
34. Preux, P. M. & Druet-Cabanac, M. Epidemiology and aetiology of epilepsy in sub-Saharan Africa. *Lancet Neurol.* **4**, 21–31 (2005).
35. Benamer, H. T. & Grosset, D. G. A systematic review of the epidemiology of epilepsy in Arab countries. *Epilepsia* **50**, 2301–2304 (2009).
36. Mirkin, B. *Population Levels, Trends and Policies in the Arab Region: Challenges and Opportunities*. Arab Human Development Report [http://mait.cam.ac.uk/ET2050\\_library/docs/med/arab\\_population.pdf](http://mait.cam.ac.uk/ET2050_library/docs/med/arab_population.pdf). (United Nations Development Programme, 2010).
37. Thomas, S. V. & Nair, A. Confronting the stigma of epilepsy. *Ann. Indian Acad. Neurol.* **14**, 158–163 (2011).
38. Fleitlich-Bilyk, B. & Goodman, R. Prevalence of child and adolescent psychiatric disorders in southeast Brazil. *J. Am. Acad. Child Adolesc. Psychiatry* **43**, 727–734 (2004).
39. Duarte, C. et al. Child mental health in Latin America: present and future epidemiologic research. *Int. J. Psychiatry Med.* **33**, 203–222 (2003).
40. Rodriguez, A. et al. Is prenatal alcohol exposure related to inattention and hyperactivity symptoms in children? Disentangling the effects of social adversity. *J. Child Psychol. Psychiatry* **50**, 1073–1083 (2009).
41. World Health Organization. *Global Status Report on Alcohol and Health* [http://apps.who.int/iris/bitstream/10665/112736/1/9789240692763\\_eng.pdf](http://apps.who.int/iris/bitstream/10665/112736/1/9789240692763_eng.pdf) (WHO, 2014).



42. Borges, G. et al. Alcohol and violence in the emergency department: a regional report from the WHO collaborative study on alcohol and injuries. *Salud Publica Mex.* **50** (Suppl 1), S6–S11 (2008).
43. Massey, J. S., Meares, S., Batchelor, J. & Bryant, R. A. An exploratory study of the association of acute posttraumatic stress, depression, and pain to cognitive functioning in mild traumatic brain injury. *Neuropsychology* **29**, 530–542 (2015).
44. Burneo, J. G., Tellez-Zenteno, J. & Wiebe, S. Understanding the burden of epilepsy in Latin America: a systematic review of its prevalence and incidence. *Epilepsy Res.* **66**, 63–74 (2005).
45. Bruno, E. et al. Epilepsy and neurocysticercosis in Latin America: a systematic review and meta-analysis. *PLoS Negl. Trop. Dis.* **7**, e2480 (2013).
46. Nitrini, R. et al. Prevalence of dementia in Latin America: a collaborative study of population-based cohorts. *Int. Psychogeriatr.* **21**, 622–630 (2009).
47. Prince, M. et al. The global prevalence of dementia: a systematic review and metaanalysis. *Alzheimers Dement.* **9**, 63–75 (2013).
48. Mejia, S., Giraldo, M., Pineda, D., Ardila, A. & Lopera, F. Nongenetic factors as modifiers of the age of onset of familial Alzheimer's disease. *Int. Psychogeriatr.* **15**, 337–349 (2003).
49. Paradisi, I., Hernandez, A. & Arias, S. Huntington disease mutation in Venezuela: age of onset, haplotype analyses and geographic aggregation. *J. Hum. Genet.* **53**, 127–135 (2008).
50. Pastor, P. et al. Apolipoprotein E4 modifies Alzheimer's disease onset in an E280A PS1 kindred. *Ann. Neurol.* **54**, 163–169 (2003).
51. Feigin, V. L. et al. Global and regional burden of stroke during 1990–2010: findings from the Global Burden of Disease Study 2010. *Lancet* **383**, 245–254 (2014).
52. Mac, T. L. et al. Epidemiology, aetiology, and clinical management of epilepsy in Asia: a systematic review. *Lancet Neurol.* **6**, 533–543 (2007).
53. Alzheimer's Disease International. *World Alzheimer Report 2009: The Global Prevalence of Dementia* <http://www.alz.co.uk/research/world-report-2009> (ADI, 2009).
54. Shaji, S., Bose, S. & Verghese, A. Prevalence of dementia in an urban population in Kerala, India. *Br. J. Psychiatry* **186**, 136–140 (2005).
55. Mehndiratta, M. M., Khan, M., Mehndiratta, P. & Wasay, M. Stroke in Asia: geographical variations and temporal trends. *J. Neurol. Neurosurg. Psychiatry* **85**, 1308–1312 (2014).
56. Kulshreshtha, A., Anderson, L. M., Goyal, A. & Keenan, N. L. Stroke in South Asia: a systematic review of epidemiologic literature from 1980 to 2010. *Neuroepidemiology* **38**, 123–129 (2012).
57. Wasay, M., Khatiri, I. A. & Kaul, S. Stroke in South Asian countries. *Nature Rev. Neurol.* **10**, 135–143 (2014).
58. Xu, G., Ma, M., Liu, X. & Hankey, G. J. Is there a stroke belt in China and why? *Stroke* **44**, 1775–1783 (2013).
59. Mackay, J., Ritthiphakdee, B. & Reddy, K. S. Tobacco control in Asia. *Lancet* **381**, 1581–1587 (2013).
60. Dans, A. et al. The rise of chronic non-communicable diseases in southeast Asia: time for action. *Lancet* **377**, 680–689 (2011).
61. Uthman, O. A. et al. Increasing the value of health research in the WHO African Region beyond 2015 – reflecting on the past, celebrating the present and building the future: a bibliometric analysis. *BMJ Open* **5**, e006340 (2015).
62. Razzouk, D. et al. Scarcity and inequity of mental health research resources in low-and-middle income countries: a global survey. *Health Policy* **94**, 211–220 (2010).
63. Sharan, P. et al. Mental health research priorities in low- and middle-income countries of Africa, Asia, Latin America and the Caribbean. *Br. J. Psychiatry* **195**, 354–363 (2009).
64. Okasha, A., Karam, E. & Okasha, T. Mental health services in the Arab world. *World Psychiatry* **11**, 52–54 (2012).
65. Goel, D. S., Agarwal, S. P., Ichhapujari, R. L. & Shrivastava, S. In *Mental Health: An Indian Perspective, 1946–2003* (eds S.P. Agarwal et al.) 3–24 (Directorate General of Health Services/Ministry of Health and Family Welfare, 2004).
66. Seedat, S. et al. Twelve-month treatment of psychiatric disorders in the South African Stress and Health Study (World Mental Health Survey Initiative). *Soc. Psychiatry Psychiatr. Epidemiol.* **43**, 889–897 (2008).
67. World Health Organization. *Atlas: Country Resources for Neurological Disorders* [http://www.who.int/mental\\_health/neurology/neurogy\\_atlas\\_lr.pdf](http://www.who.int/mental_health/neurology/neurogy_atlas_lr.pdf) (WHO, 2004).
68. Saxena, S., Paraje, G., Sharan, P., Karam, G. & Sadana, R. The 10/90 divide in mental health research: trends over a 10-year period. *Br. J. Psychiatry* **188**, 81–82 (2006).
69. Pang, T., Lansang, M. A. & Haines, A. Brain drain and health professionals. *Br. Med. J.* **324**, 499–500 (2002).
70. Padma, T. V. Developing countries: the outcomes paradox. *Nature* **508**, S14–15 (2014).
71. Dorsey, E. R. et al. Projected number of people with Parkinson disease in the most populous nations, 2005 through 2030. *Neurology* **68**, 384–386 (2007).
72. Paddick, S. M. et al. Dementia prevalence estimates in sub-Saharan Africa: comparison of two diagnostic criteria. *Glob. Health Action* **6**, 19646 (2013).
73. Yang, G. et al. Rapid health transition in China, 1990–2010: findings from the Global Burden of Disease Study 2010. *Lancet* **381**, 1987–2015 (2013).
74. Irving, J. A. & Segal, Z. V. Mindfulness-based cognitive therapy: current status and future applications. *Sante Ment. Que.* **38**, 65–82 (2013).
75. Kabat-Zinn, J. et al. Effectiveness of a meditation-based stress reduction program in the treatment of anxiety disorders. *Am. J. Psychiatry* **149**, 936–943 (1992).
76. Linehan, M. M. Dialectical behavior therapy for borderline personality disorder. Theory and method. *Bull. Menninger. Clin.* **51**, 261–276 (1987).
77. Manjunath, R. B., Varambally, S., Thirthalli, J., Basavaraddi, I. V. & Gangadhar, B. N. Efficacy of yoga as an add-on treatment for in-patients with functional psychotic disorder. *Indian J. Psychiatry* **55**, S374–S378 (2013).
78. Rao, N. P., Varambally, S. & Gangadhar, B. N. Yoga school of thought and psychiatry: Therapeutic potential. *Indian J. Psychiatry* **55**, S145–S149 (2013).
79. Knudsen, E. I., Heckman, J. J., Cameron, J. L. & Shonkoff, J. P. Economic, neurobiological, and behavioral perspectives on building America's future workforce. *Proc. Natl Acad. Sci. USA* **103**, 10155–10162 (2006).
80. Bain, L. E. et al. Malnutrition in Sub-Saharan Africa: burden, causes and prospects. *Pan. Afr. Med. J.* **15**, 120 (2013).
81. Motadi, S. A., Mbhenyane, X. G., Mbhatsani, H. V., Mabapa, N. S. & Mamabolo, R. L. Prevalence of iron and zinc deficiencies among preschool children ages 3 to 5 y in Vhembe district, Limpopo province, South Africa. *Nutrition* **31**, 452–458 (2015).
82. Said-Mohamed, R., Micklesfield, L. K., Pettifor, J. M. & Norris, S. A. Has the prevalence of stunting in South African children changed in 40 years? A systematic review. *BMC Public Health* **15**, 534 (2015).
83. Black, R. E. et al. Maternal and child undernutrition and overweight in low-income and middle-income countries. *Lancet* **382**, 427–451 (2013).
84. Kitsao-Wekulo, P. et al. Nutrition as an important mediator of the impact of background variables on outcome in middle childhood. *Front. Hum. Neurosci.* **7**, 713 (2013).
85. Nzwalu, H. & Cliff, J. Konzo: From poverty, cassava, and cyanogen intake to toxic-nutritional neurological disease. *PLoS Negl. Trop. Dis.* **5**, e1051 (2011).
86. Woldeamanuel, Y. W., Hassan, A. & Zenebe, G. Neuroleptism: two Ethiopian case reports and review of the literature. *J. Neurol.* **259**, 1263–1268 (2012).
87. United Nations Office on Drugs and Crime. *World Drug Report* [https://www.unodc.org/documents/wdr2014/World\\_Drug\\_Report\\_2014\\_web.pdf](https://www.unodc.org/documents/wdr2014/World_Drug_Report_2014_web.pdf). (UN, 2014).
88. Watt, M. H. et al. The impact of methamphetamine (“tik”) on a peri-urban community in Cape Town, South Africa. *Int. J. Drug Policy* **25**, 219–225 (2014).
89. Schuurman, N. et al. Intentional injury and violence in Cape Town, South Africa: an epidemiological analysis of trauma admissions data. *Glob. Health Action* **8**, 27016 (2015).
90. Hoffman, R. & al'Absi, M. Working memory and speed of information processing in chronic khat users: preliminary findings. *Eur. Addict. Res.* **19**, 1–6 (2013).
91. Duggan, M. B. Epilepsy and its effects on children and families in rural Uganda. *Afr. Health Sci.* **13**, 613–623 (2013).
92. Owolabi, M. O. et al. The burden of stroke in Africa: a glance at the present and a glimpse into the future. *Cardiovasc. J. Afr.* **26**, S27–S38 (2015).
93. Lekoubou, A., Echouffo-Tcheugui, J. B. & Kengne, A. P. Epidemiology of neurodegenerative diseases in sub-Saharan Africa: a systematic review. *BMC Public Health* **14**, 653 (2014).
94. Li, D., Zhang, D. J., Shao, J. J., Qi, X. D. & Tian, L. A meta-analysis of the prevalence of depressive symptoms in Chinese older adults. *Arch. Gerontol. Geriatr.* **58**, 1–9 (2014).
95. Ma, X., Li, S. R. & Xiang, Y. Q. An epidemiological survey on depressive disorder in Beijing Area. *Chinese J. Psychiatry* **40**, 100–103 (2007).
96. Malhotra, S. & Patra, B. N. Prevalence of child and adolescent psychiatric disorders in India: a systematic review and meta-analysis. *Child Adolesc. Psychiatry Ment. Health* **8**, 22 (2014).
97. Saxena, S., Thornicroft, G., Knapp, M. & Whiteford, H. Resources for mental health: scarcity, inequity, and inefficiency. *Lancet* **370**, 878–889 (2007).
98. Patel, V. The need for treatment evidence for common mental disorders in developing countries. *Psychol. Med.* **30**, 743–746 (2000).
99. Subodh, B. N., Avasthi, A. & Chakrabarti, S. Psychosocial impact of dysthymia: a study among married patients. *J. Affect. Disord.* **109**, 199–204 (2008).
100. Chan, K. Y. et al. Epidemiology of Alzheimer's disease and other forms of dementia in China, 1990–2010: a systematic review and analysis. *Lancet* **381**, 2016–2023 (2013).
101. Jitapunkul, S., Kunanusont, C., Phoolcharoen, W. & Suriyawongpaisal, P. Prevalence estimation of dementia among Thai elderly: a national survey. *J. Med. Assoc. Thai* **84**, 461–467 (2011).
102. Shaji, K. S. et al. *The Dementia India Report: Prevalence, Impact, Costs and Services for Dementia* [http://www.alzheimer.org.in/dementia\\_2010.pdf](http://www.alzheimer.org.in/dementia_2010.pdf). (Alzheimer's and Related Disorders Society of India, 2010).
103. Wang, G. et al. Economic impact of dementia in developing countries: an evaluation of Alzheimer-type dementia in Shanghai, China. *J. Alzheimers Dis.* **15**, 109–115 (2008).
104. Liu, L., Wang, D., Wong, K. S. & Wang, Y. Stroke and stroke care in China: huge burden, significant workload, and a national priority. *Stroke* **42**, 3651–3654 (2011).
105. Alarcon, R. D. Mental health and mental health care in Latin America. *World Psychiatry* **2**, 54–56 (2003).
106. Hyder, A. A., Wunderlich, C. A., Puvanachandra, P., Gururaj, G. & Kobusingye, O. C. The impact of traumatic brain injuries: a global perspective. *Neurorehabilitation* **22**, 341–353 (2007).
107. D'Souza, M. S. & Markou, A. Neuronal mechanisms underlying development of nicotine dependence: implications for novel smoking-cessation treatments. *Addict. Sci. Clin. Pract.* **6**, 4–16 (2011).
108. Pyne, H. H., Claeson, M. & Correia, M. *Gender Dimensions of Alcohol Consumption and Alcohol-Related Problems in Latin America and the Caribbean. World Bank Discussion paper; no. WDP 433* [http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP/IB/2005/04/25/000112742\\_20050425144138/Rendered/PDF/wdp435.pdf](http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP/IB/2005/04/25/000112742_20050425144138/Rendered/PDF/wdp435.pdf). (World Bank, 2002).
109. Champagne, B. M. et al. Tobacco smoking in seven Latin American cities: the CAR-MELA study. *Tob. Control* **19**, 457–462 (2010).
110. Barreto, S. M. et al. Epidemiology in Latin America and the Caribbean: current situation and challenges. *Int. J. Epidemiol.* **41**, 557–571 (2012).

111. Hyder, A. A. et al. Global childhood unintentional injury surveillance in four cities in developing countries: a pilot study. *Bull World Health Organ.* **87**, 345–352 (2009).
112. Puvanachandra, P. & Hyder, A. A. Traumatic brain injury in Latin America and the Caribbean: a call for research. *Salud Publica Mex.* **50 Suppl 1**, S3–S5 (2008).
113. Travers, K. U., Pokora, T. D., Cadarette, S. M. & Mould, J. F. Major depressive disorder in Africa and the Middle East: a systematic literature review. *Expert Rev. Pharmacoecon. Outcomes Res.* **13**, 613–630 (2013).
114. World Health Organization. *Iraq Family Health Survey Report 2006/7* [http://www.who.int/mediacentre/news/releases/2008/pr02/2008\\_iraq\\_family\\_health\\_survey\\_report.pdf](http://www.who.int/mediacentre/news/releases/2008/pr02/2008_iraq_family_health_survey_report.pdf) (WHO, 2007).
115. Lavi, T. & Solomon, Z. Palestinian youth of the Intifada: PTSD and future orientation. *J Am. Acad. Child Adolesc. Psychiatry* **44**, 1176–1183 (2005).
116. Bassiony, M. M. et al. Adolescent tramadol use and abuse in Egypt. *Am. J. Drug Alcohol Abuse* **41**, 206–211 (2015).
117. Khattab, A. et al. Smoking habits in the Middle East and North Africa: results of the BREATHE study. *Respir. Med.* **106** (Suppl 2), S16–S24 (2012).
118. El-Zaemey, S., Heyworth, J. & Fritschi, L. Qat consumption among women living in Yemen. *Int. J. Occup. Environ. Med.* **5**, 109–111 (2014).
119. Momtazi, S. & Rawson, R. Substance abuse among Iranian high school students. *Curr. Opin. Psychiatry* **23**, 221–226 (2010).

#### ACKNOWLEDGMENTS

The authors thank N. Rao at the Centre for Neuroscience, Indian Institute of Science for his help with the manuscript.

#### COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests. Financial support for publication has been provided by the Fogarty International Center.

#### ADDITIONAL INFORMATION



This work is licensed under the Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0>

REVIEW **OPEN**

# Building global capacity for brain and nervous system disorders research

Linda B. Cottler<sup>1</sup>, Joseph Zunt<sup>2</sup>, Bahr Weiss<sup>3</sup>, Ayesha Kamran Kamal<sup>4</sup> & Krishna Vaddiparti<sup>1</sup>

The global burden of neurological, neuropsychiatric, substance-use and neurodevelopmental disorders in low- and middle-income countries is worsened, not only by the lack of targeted research funding, but also by the lack of relevant in-country research capacity. Such capacity, from the individual to the national level, is necessary to address the problems within a local context. As for many health issues in these countries, the ability to address this burden requires development of research infrastructure and a trained cadre of clinicians and scientists who can ask the right questions, and conduct, manage, apply and disseminate research for practice and policy. This Review describes some of the evolving issues, knowledge and programmes focused on building research capacity in low- and middle-income countries in general and for brain and nervous system disorders in particular.

*Nature* 527, S207–S213 (19 Month 2015), DOI: 10.1038/nature16037

This article has not been written or reviewed by *Nature* editors. *Nature* accepts no responsibility for the accuracy of the information provided.

**D**espite the current global burden of neurological, mental health (neuropsychiatric), developmental and substance use (NMDs) disorders, which is projected to increase, there is a lack of well-trained clinicians and scientists who focus on brain and nervous system disorders research in low- and middle-income countries (LMICs)<sup>1–6</sup> (see page S151). This workforce deficiency limits advances in research that can lead to new and improved interventions for those who are living with brain and nervous-system disorders. Notably, there are 200 times more neurologists per capita, and up to 160 times more psychiatrists per capita, in high-income countries than there are in LMICs<sup>7,8</sup>. Brain disorders involve central nervous system (CNS) functioning, making things even more challenging. Key symptoms may involve both internalizing and externalizing behaviour. Internalizing behaviour may be stigmatized and externalizing behaviour may be negatively stereotyped; both may be difficult to treat and associated with poor prognosis. The stigma associated with NMDs disorders applies to both the patients and the clinician researchers who treat them, creating an additional barrier to building and sustaining research capacity.

Thus, for health research in LMICs there is an urgent need to build new, and strengthen existing, individual, institutional and country-wide research capabilities. This Review identifies certain key characteristics of capacity building, and the challenges and lessons learned based on the literature<sup>9,10</sup> and our own experience.

## BUILDING AND STRENGTHENING RESEARCH CAPACITY

Research capacity building is a systematic, purposeful and goal-oriented effort to strengthen human resources and infrastructure to enable local scientists and institutions to become independent and responsive to existing and emerging health needs and threats<sup>9–13</sup>. To be sustainable and effective (and to address research-training sustainability

concerns) a framework must be created so that research capacity is strengthened and woven together at the individual, institutional and national levels. To create research opportunity and a career pipeline, there needs to be a simultaneous focus on frameworks, goals and opportunities. Embedding research into health systems requires a process that involves competent scientists and a supportive environment that enables research communities to flourish as they use new research tools that contribute to improving the health of the population<sup>12</sup>. This iterative process allows research to be responsive to population needs, and policies and practices to be responsive to research findings.

## INDIVIDUAL LEVEL

At the core of capacity building is the training and mentoring of individuals to design and conduct research; to create or adapt research tools that are relevant to brain-disorders research; to form collaborations with scientists in their institution, elsewhere in the country, and internationally; and eventually to serve as mentors themselves for the next generation of scientists (most effectively if they are within their home countries or region). Increasingly, researchers also need training on how to interact effectively with policy and programme implementers to ensure that they and their practices are adapted to the practices and policies locally.

Capacity building at the individual level begins with mentors who counsel, provide career guidance, and advise on the teaching and sharing of ethical principles that instil integrity in research and care. Research-mentoring strategies and systems are context-driven, and often use apprenticeship or hands-on models, whereby individuals learn by doing. When little in-country research expertise and capacity exists, training may need to take place initially in a higher resource country. However, the goal should always be to have the research training in the context of the trainees' home institution.

<sup>1</sup>Department of Epidemiology, College of Public Health and Health Professions and College of Medicine, University of Florida, Gainesville, Florida, USA. <sup>2</sup>Department of Neurology, University of Washington, Seattle, Washington, USA. <sup>3</sup>Department of Psychology and Human Development, Vanderbilt University, Nashville, Tennessee, USA. <sup>4</sup>Stroke Service, Section of Neurology, Department of Medicine, The International Cerebrovascular Translational Clinical Research Training Program (Fogarty International Center, NIH) Aga Khan University, Karachi, Pakistan. Correspondence should be addressed to L. B. C. e-mail: lbcottler@ufl.edu.



## BOX 1 | BIOMEDICAL AND BIOBEHAVIOURAL RESEARCH ADMINISTRATOR DEVELOPMENT PROGRAMME

The National Institutes of Health (NIH)/National Institute of Child Health and Human Development (NICHD) in collaboration with other NIH Institutions and Centers has made available the Biomedical/Biobehavioral Research Administrator Development (BRAD; <http://www.nichd.nih.gov/about/org/od/ohe/brad/Pages/overview.aspx>) programme to establish new or strengthen existing offices of sponsored programmes (OSPs) or similar entities in low- and middle-income countries (LMICs). Enhanced OSPs in non-research intensive institutions are essential for the development of enabling and supportive environments in which faculty can develop robust research programmes and provide research experiences for students.

### BRAD objectives

From a programme perspective, comprehensive and effective research administration is the bridge between research projects and a sustainable research enterprise at LMIC institutions. Accordingly, the BRAD programme objectives are to:

- Encourage and support continuous professional development of OSPs, research administrators and grants managers at all levels.
- Increase the effectiveness and productivity of OSPs (or similar entities) by promoting the use of best practices in research administration.
- Promote OSP sustainability by identifying and addressing barriers to research and by supporting targeted faculty professional development that focuses on increasing competitiveness in obtaining external research support.

In terms of systems, scientists initially educated in LMICs then trained in high-income countries say that they are accustomed to certain ways of providing and receiving feedback, and interacting within hierarchical relationships. When they return to their home academic environments, ideally they will adopt a hybrid mentoring style, combining positive attributes from both the country where they were trained and their home country, of which they have an awareness of local nuances, regarding local customs, politics and bureaucracy.

Questioning and vigorous debate are integral to the scientific process, but young investigators from cultures that emphasize deference in hierarchical relationships may experience a conflict between these two values when they return home.

In terms of research and clinical practice, performing a lumbar puncture to obtain cerebrospinal fluid (CSF), a requirement for the diagnosis of most CNS infections and some degenerative or developmental abnormalities, brain banking and autopsy can all be met with reluctance in LMICs. Extraction of fluids, tissue or organs for donation or banking for research purposes is associated with significant sociocultural barriers in some cultures. Several factors related to religion and culture, and issues related to distrust of the medical system, misunderstandings about religious stances and ignorance often complicate the process. Requests for brain or other organ banking for research purposes could raise concerns that agreeing to donation would discourage doctors from treatment to save lives among relatives or that consenting to banking would result in premature removal of their or their relative's organs.

To build capacity for these endeavours, mentors must be willing to not just advocate for these techniques, but also to address cultural barriers that may affect policies, as well as the discomfort of trainees who may not have experience of approaching relatives and patients about these procedures.

The National Institutes of Health/Fogarty International Center Global Brain, NIDA International Fellowships and other programmes

(see Supplementary Table) are designed to help catalyse research capacity development at an individual level (both as research training opportunities for young investigators, and as research pipeline opportunities for more advanced LMIC investigators) to help prevent the loss of crucial talent and expertise.

## INSTITUTIONAL LEVEL

Institutional capacity building is the administrative foundation and is essential for establishing and sustaining initiatives intended to realize its vision<sup>14</sup>. Research infrastructure includes job positions that provide protected time for research, as well as robust laboratories and clinical spaces where diagnosis, treatment and research can be conducted. Research into brain disorders, especially stroke, CNS infections, trauma and neurodegenerative conditions requires the technology to assess structural neurological abnormalities; for example, computerized tomography or magnetic resonance imaging may be non-existent or prohibitively expensive in many LMIC settings. Although research-training grants typically provide the funding necessary to train new scientists and the equipment to increase laboratory capacity, larger infrastructure capacity-building endeavours, such as acquiring high-cost diagnostic neuroimaging or laboratory equipment, or constructing new laboratories, clinics or classrooms, require the financial commitment of institutions with support from funders (ideally, and eventually, at the country level for maximum sustainability).

Research and grants administration are crucial to the sustainability of research programmes within any institution, but good administrators can enhance the development of research capacity in resource-challenged institutions (Box 1).

## Networks

Research and research-training networks enrich the research environment and build capacity by increasing collaborations and partnerships; expanding institutional perspectives from local and national levels to regional and perhaps the global level; and facilitating ideas exchange, dialogue and universal or standardized protocols for brain research<sup>15</sup>. Such networks are most effective when they attract not only individual scientists and academic institutions, but also non-governmental organizations (NGOs), corporations, policymakers, and/or philanthropists, to sustain and embed the research enterprise within a country that is focused on a health issue, such as NMDs disorders. One example is the neuroscience promotion association APRONES. This association was established by a group of neurologists to share knowledge of diseases of the nervous system in LMICs. Members are from Africa, Europe and the United States. The association encourages collaborative studies around the world, while building networks and ultimately research capacity<sup>16</sup>.

## NATIONAL LEVEL

True sustainability of research capacity and its application requires a national commitment to the research enterprise and implementation of research results at the policy level, as well as a continuing dialogue between health practitioners, policymakers and researchers. However, often it is not until research and research-training networks are established that local government and NGOs recognize the benefit of talent and training to system-wide improvements and national human-resource development<sup>12</sup>. At that point, they take actions to sustain it.

According to the WHO<sup>9</sup>, work in support of the ethical review and public accountability of research is not keeping pace with best practices. Opportunities to create a shared framework for storing and sharing research data, tools and materials, have not been met with the same energy in the area of health as they have in other scientific fields. Furthermore, policymakers rarely understand research priorities or use evidence to inform their decisions.

Without country-level planning and action, along with guidance documents, health research in LMICs may be influenced more by the demands of foreign funders' research and infrastructure interests than

Table 1 | Approaches to building research capacity

Category	Activity	General and specific needs	Anticipated benefits	Specific approaches
Human capacity	<ul style="list-style-type: none"> <li>Increase the number of clinician researchers</li> </ul>	<ul style="list-style-type: none"> <li>Increased number of physicians and allied health professionals in research benefits all neuroscience researchers, including neurologists, neurosurgeons, infectious disease specialists, psychiatrists and other mental health practitioners</li> <li>As the incidence of neurological conditions increases, so will the need for more trained neurologists involved in research</li> </ul>	<ul style="list-style-type: none"> <li>Practical experience and opportunities for future training<sup>12</sup></li> </ul>	<ul style="list-style-type: none"> <li>Create protected time and funding for research</li> <li>Decrease brain drain by investing in research and jobs in neurological areas</li> <li>Research methodology training during graduate and post-graduate medical training</li> </ul>
	<ul style="list-style-type: none"> <li>Increase research capacity of clinicians and researchers through workshops and short courses; and advanced degrees in public health (for example, epidemiology, biostatistics, clinical trials, health services and implementation science), clinical and basic science research</li> <li>Sub-specialized training on specific skills related to nervous-system disorders</li> </ul>	<ul style="list-style-type: none"> <li>As the neurosciences have not received as much attention from research training grants, most areas would benefit from increased funding for research training</li> <li>Health-systems research is needed, which necessitates training in bioethics, research methodology, epidemiology, clinical trials, population-based methodology and intervention studies</li> <li>Specific areas of neuroscience with unique needs, include mental health for which health-services research is crucial to increase the capacity of care services</li> <li>Many countries do not have the ability to diagnose neurogenetic conditions, and cannot provide genetic counselling or treatment</li> </ul>	<ul style="list-style-type: none"> <li>Address the shared burdens of common conditions, including neurodegenerative disorders, stroke and epilepsy</li> <li>Development of multidisciplinary teams of health-care professionals to improve prevention, pre-hospital care, and clinical care in neuroscience, such as trauma, mental health or neurogenetics</li> </ul>	<ul style="list-style-type: none"> <li>Multidisciplinary training and research</li> <li>The Wellcome trust-DBT India alliance fellowship for clinicians and research scientists</li> <li>Innovation in science pursuit for inspired research programme</li> <li>Initiative in neuroclinical research education</li> <li>The African Brain Mapping and Therapeutics Initiative, which advances neuroscience research in Africa by promoting global partnerships for brain-disease prevention and treatment</li> </ul>
	<ul style="list-style-type: none"> <li>Institutionalization of mentorship training</li> </ul>	<ul style="list-style-type: none"> <li>Outstanding mentoring is a prerequisite of any successful research-training programme</li> <li>Mentors must be expert in particular areas of research, such as cognitive assessment scales for the study of dementia-associated conditions, or seizure management for studies of epilepsy</li> </ul>	<ul style="list-style-type: none"> <li>Capacity for conducting neuroscience research will increase as trainees move into positions where they will start mentoring subsequent generations of trainees</li> <li>Increasing numbers of scientists and the development of research teams, research culture and an increase in scientific literature and novel research</li> </ul>	<ul style="list-style-type: none"> <li>Wide spread mentorship training (for example, through programmes such as those of the NIH Fogarty International Center <a href="http://www.fic.nih.gov/">http://www.fic.nih.gov/</a>)</li> </ul>
Infrastructure and tools	<ul style="list-style-type: none"> <li>Neuroimaging (for example, computerized tomography or magnetic resonance imaging)</li> </ul>	<ul style="list-style-type: none"> <li>Most neurological conditions require neuroimaging to confirm a diagnosis, disease stage or to monitor progress</li> </ul>	<ul style="list-style-type: none"> <li>Increased availability of neuroimaging will lead to better definition of the burden of many neurological conditions, such as stroke, CNS infections, developmental, degenerative and genetic disorders, and trauma</li> </ul>	<ul style="list-style-type: none"> <li>Mentoring in the tools needed through flexible research and research-training programmes (for example through the Fogarty Global Brain programme)</li> </ul>
	<ul style="list-style-type: none"> <li>Genomic sequencing to detect SNPs in GWAS</li> </ul>	<ul style="list-style-type: none"> <li>GWAS are used to identify genetic variations (SNPs) associated with neurological and psychological disorders, including addiction</li> </ul>	<ul style="list-style-type: none"> <li>Detection of specific genes through GWAS can lead to a better understanding of the functional mechanisms that are biologically important in disease pathogenesis and, ultimately, to better treatments for neurological diseases</li> </ul>	<ul style="list-style-type: none"> <li>The US National Center for Biotechnology Information has developed the Database of Genotype and Phenotype, where genetic sequencing information can be deposited and accessed</li> </ul>
	<ul style="list-style-type: none"> <li>Increased laboratory capacity</li> </ul>	<ul style="list-style-type: none"> <li>Most studies of neurological diseases require at least a basic laboratory to process blood, cerebrospinal fluid or other human samples</li> <li>With increasing complexity of studies, additional equipment is needed, such as polymerase chain reaction for detecting infectious pathogens or biosensors to detect environmental toxins</li> </ul>	<ul style="list-style-type: none"> <li>Many technologies introduced to increase laboratory capacity are also useful for diagnosing non-neurological diseases</li> </ul>	<ul style="list-style-type: none"> <li>Enhancement of laboratory capacity often requires the upgrade of electrical systems, and with larger laboratories may also require installation of air-conditioning systems</li> </ul>
	<ul style="list-style-type: none"> <li>Culturally appropriate assessment and screening tools</li> </ul>	<ul style="list-style-type: none"> <li>WHO and NIH databank of valid and reliable assessments for young people and adults. Each has an armamentarium of tools that are culturally appropriate</li> <li>Stigma, social and health disparities are more common with disorders such as epilepsy and schizophrenia</li> </ul>	<ul style="list-style-type: none"> <li>Disorders such as epilepsy and schizophrenia would benefit from increased recognition of barriers identified through culturally appropriate screening tools</li> </ul>	<ul style="list-style-type: none"> <li>NIH and WHO promote scientific discovery, and shared resources, that allow for data harmonization across many programmes</li> <li>Network meetings with special interest groups value the use of unified concepts of addiction and mental health, from DSM to ICD classifications <b>cont.</b></li> </ul>

by the health priorities of the host country<sup>17</sup>. In LMICs in general, and Africa specifically, increasing the value of health research requires evidence-informed actions to be taken by relevant authorities to ensure that health research is conspicuous in development agendas. It also requires defining, financing and monitoring a clear national plan for a future research enterprise focused on health. To achieve these goals, policymakers and public health and research-funding institutions can use principles adapted from the WHO Strategy on Research for Health<sup>9</sup> as a guide. These provide the overall framework for research capacity and include reinforcing the research culture and organization;

focusing research on key health challenges by setting priorities; strengthening national health research systems and building capacity; encouraging good research practice (setting standards) and consolidating links between health research and action (translation and evidence-based implementation).

Needs and opportunities for building and strengthening capacity for brain-disorders research are shown in Table 1. Although not exhaustive, they outline specific valuable approaches that can be used by high-income country and LMIC collaborators.

The Supplementary Table includes some organizations that are

Category	Activity	General and specific needs	Anticipated benefits	Specific approaches
<b>Technology</b>	<ul style="list-style-type: none"> <li>Incorporation of emerging POC diagnostics from both the development of cross-cultural tools to the use of the tools</li> </ul>	<ul style="list-style-type: none"> <li>POC diagnostics could permit rapid diagnosis of many neurological infections in the field, resulting in improved recognition and treatment</li> <li>POC diagnostics can be used to non-invasively monitor seizures, cerebral blood flow or intracranial pressure, but are not widely available for use in LMIC settings</li> </ul>	<ul style="list-style-type: none"> <li>Miniaturization of diagnostic technologies for genomics, infectious agents and environmental markers will enable a better understanding of gene–environment interactions and lead to new therapeutic approaches</li> </ul>	<ul style="list-style-type: none"> <li>Better sharing of data and instrumentation is needed, as well as collaborative grants</li> </ul>
	<ul style="list-style-type: none"> <li>Access to electronic scientific literature</li> </ul>	<ul style="list-style-type: none"> <li>Common to all research is the need for understanding past and current scientific literature</li> </ul>	<ul style="list-style-type: none"> <li>Improved access to electronic scientific literature should lead to more scientifically sound research and often leads to the creation of journal clubs, which in turn strengthens the culture of research</li> </ul>	<ul style="list-style-type: none"> <li>The HINARI Access to Research in Health programme provides free or low cost access to 200 neuroscience journals for not-for-profits in LMICs, but the top ranking 50 journals are not available</li> <li>Open access journals are available to everyone</li> </ul>
	<ul style="list-style-type: none"> <li>Introduction of mHealth technologies and e-learning strategies</li> </ul>	<ul style="list-style-type: none"> <li>Adoption of mobile technologies for surveillance, assessments and treatment are particularly needed in LMICs where cell phone ownership is rising rapidly, but access to conventional health care and health-care providers is limited</li> <li>Modular Internet-based curricula can be adapted for training for advancing neuroscience research</li> <li>Low-tech clinical simulation training emphasizing early life-saving interventions and procedures</li> </ul>	<ul style="list-style-type: none"> <li>Reaching patients with disabling neurological conditions using cell phones may prove easier than conventional methods for providing health information</li> </ul>	<ul style="list-style-type: none"> <li>Share resources</li> <li>Offer classes for students at reduced cost</li> </ul>
	<ul style="list-style-type: none"> <li>Increase Internet capacity</li> </ul>	<ul style="list-style-type: none"> <li>Adapt information and communication technologies to support research and research-training programmes</li> </ul>	<ul style="list-style-type: none"> <li>Video conferencing for direct communication between mentors, colleagues and training in diverse settings</li> </ul>	<ul style="list-style-type: none"> <li>Communication technologies include Skype, GoToMeeting, AdobeConnect, WhatsApp, Polycom and WebEx</li> </ul>
<b>Funding</b>	<ul style="list-style-type: none"> <li>Pilot awards for LMIC researchers</li> </ul>	<ul style="list-style-type: none"> <li>Funding for research in LMICs is limited, but funding for neurological disease research is even more so</li> </ul>	<ul style="list-style-type: none"> <li>Providing funding to support pilot studies to LMIC colleagues and trainees should lead to increased research relevant to the LMIC setting and provide pilot data on which larger grant applications could be developed</li> </ul>	<ul style="list-style-type: none"> <li>Mentor LMIC partners through application processes.</li> </ul>
	<ul style="list-style-type: none"> <li>Increase governmental funding for research through universities and research institutions</li> </ul>	<ul style="list-style-type: none"> <li>Funds from national and international NGOs can increase research opportunities</li> </ul>	<ul style="list-style-type: none"> <li>Collaboration with foreign partners provides new research opportunities and support</li> </ul>	<ul style="list-style-type: none"> <li>PEPFAR, UNAIDS, WHO, and the Bill and Melinda Gates Foundation have made drugs and services significantly more accessible</li> </ul>
	<ul style="list-style-type: none"> <li>Research frameworks to support the implementation of the outcomes of well-designed studies relevant to neurological diseases</li> </ul>	<ul style="list-style-type: none"> <li>Evidence-based public health strategies to incorporate child neurodisability screening, clinical evaluation and rehabilitation packages into the health-care system</li> </ul>	<ul style="list-style-type: none"> <li>Maternal health programmes that work closely with early childhood programmes could ensure optimal pregnancy outcomes and develop effective interventions to enhance child development</li> </ul>	
	<ul style="list-style-type: none"> <li>Capacity building in translational science and knowledge management</li> <li>Learn to package the evidence in a format more accessible to policymakers in LMICs</li> </ul>	<ul style="list-style-type: none"> <li>LMIC partners are asking for translational science training</li> </ul>	<ul style="list-style-type: none"> <li>Influence policymakers to redirect budget priorities to address brain disorders and their research</li> <li>Involve community partnerships</li> </ul>	<ul style="list-style-type: none"> <li>Involvement of the community, private sector and research sponsors during project planning establishes priorities, identifies research needs within the community, and identifies resources<sup>34</sup></li> </ul>

CNS, central nervous system; DSM, Diagnostic and Statistical Manual for Mental Disorder; GWAS, genome-wide association studies; ICD, International Classification of Diseases; LMIC, low- and middle-income countries; NGOs, non-governmental organizations; POC, point-of-care; SNP, single nucleotide polymorphisms; WHO, World Health Organization.

investing in research-capacity building for brain disorders. Some long-term examples of programmes specifically focused on building a pipeline that stretches from individual to institutional to national research capacity levels for nervous system diseases and disorders in LMICs are shown in Box 2. These programmes include the US NIH/Fogarty coordinated Global Brain and Nervous System Diseases and Disorders Across the Lifespan Research Program and several Fogarty centre sponsored institutional research training programmes (focusing on masters, PhD and postdoctoral level training for LMIC investigators; Supplementary Table).

## ADDRESSING CHALLENGES TO CAPACITY BUILDING

The framework to build the individual, institutional and national capacity described requires principles of engagement, much like those for community engagement<sup>18</sup>. Those most relevant are the seven principles from the ESSENCE good practice document series<sup>10</sup> (Box 3). These core principles serve as a useful guide for funding agencies, the scientific community and academic institutions on how to move forward as they identify priorities, develop goals and objectives, design programmes and establish partnerships. They are also useful principles

to address various challenges in collaborations within and between countries and cultures, such as human, infrastructure, technological and ethical challenges.

## Human capacity challenges

Capacity building across countries and cultures for brain-disorders research is inherently a ‘messy’ processes when we consider the scope of research with global partners across completely different time zones, infrastructures, cultural norms, expectations and organizational research capacities. Differences in language and expression of research ideas can lead to confusion and misunderstanding. When choosing terminology for assessments on depression, for example, well-developed Western assessments use the words ‘feeling blue’ to indicate feelings of sadness. To discuss mania, the term ‘high’ might be used. These terms are idiomatic and do not translate well in many languages. Another example is the need for translators and interpreters in different sites within countries where multiple languages and dialects are spoken.

Research training includes emphasizing flexibility to address such cultural challenges to research and working within local and regional



## BOX 2 | ANATOMY OF A GLOBAL BRAIN RESEARCH FUNDING PROGRAMME

Achievements of the National Institutes of Health/Fogarty International Centre coordinated global brain research programme (<http://www.fic.nih.gov/About/Staff/Policy-Planning-Evaluation/Pages/fogarty-program-evaluation-brain-disorders.aspx>).

The programme supports collaborative empirical research and capacity building on brain and nervous-system diseases and disorders identified by the applicants as relevant public health challenges in their low- and middle-income countries (LMICs).

### Programme's achievements

Research conducted over 10 years in 45 LMICs, most of which are in sub-Saharan Africa, Latin America and the Caribbean.

Topics were across the spectrum, from mental health and substance use, to peripheral nervous system trauma and gene–environment interactions.

During the first 10 years of the programme, participants published 435 peer-reviewed articles in 249 unique journals, as well as 14 books or chapters.

Grantees also produced unique tools for clinical assessment in the LMIC context, developed and evaluated new interventions, and identified novel laboratory tools or methods.

Almost half of the projects supported training for people, who were not primary collaborators, in LMICs. The programme supported in-depth instruction for at least 138 scientists, for an average of 23 months.

Projects included training or mentoring at the LMIC (or sometimes a high-income country) site, in skills, methods or procedures that are essential to research, including workshops on specific topics, or clinical or research skills.

Achieved mandatory training in research ethics, which built and sustained capacity in research ethics at most sites.

norms (the ESSENCE principles shown in Box 3 can help). For example, when conducting a study on the prevalence of opiate use in Afghanistan, two female interviewers were needed to conduct interviews with the female head of the household. Cultural norms dictated that the female interviewers had to be accompanied by a male team member who would make the first contact with the household<sup>19</sup>. Designing research protocols that account for local cultural norms, while educating the high-income country and LMIC institutional review boards is necessary and builds trust and understanding between collaborators over time. Individual challenges can be resolved through open communication and sharing of expectations at the outset.

### Infrastructure capacity challenges

To conduct research, LMIC institutions require institutional review board committees, grant-management personnel, and data and document management capacities. These capacities vary widely across and within countries, but sufficient capacity in these areas is crucial to ensure fidelity to research protocols. Financial resource limitations and limited access to scientific and technical information are also key challenges.

To overcome these barriers, high-income country and LMIC partners have strengthened institutional support by setting up meetings with presidents, deans and directors of institutes to advocate for more resources and to become less reliant on outside high-income country funding. Researchers have also set up channels of communication and collaboration whereby they can help each other in the grant application

## BOX 3 | SEVEN PRINCIPLES FOR STRENGTHENING RESEARCH CAPACITY

Based on the World Health Organization–TDR ESSENCE good practice document series<sup>10</sup>

1. Network, collaborate, communicate and share experiences
2. Understand the local context and accurately evaluate existing research capacity
3. Ensure local ownership and secure active support
4. Build in monitoring, evaluation and learning from the start
5. Establish robust research governance and support structure, and promote effective leadership
6. Embed strong support, supervision and mentorship structures
7. Think long-term, be flexible and plan for continuity

process for Western-based grants.

Investigators have also succeeded in seeking permanent access to library resources through high-income country institutions for their trainees. However, more sustainable access within and across LMICs to bridge the global information divide is needed. One source is HINARI<sup>20</sup>.

### Technology capacity challenges

Information and communication technology (ICT) has become increasingly integrated into research and clinical training. ICT involves a variety of technologies, including low-cost two-way voice, picture and video communication; development of geographic information systems, which are useful for planning interventions and mapping the prevalence of neurological conditions and risk factors<sup>21,22</sup>; Internet- and mobile-phone-based health-related interventions<sup>23</sup>, and Internet- and mobile-phone-based data collection<sup>24–32</sup>.

Online courses and degree programmes that have become incorporated into most high-income country academic institutions have particular utility in LMICs where training infrastructure may be lacking or geographical barriers limit participation in conventionally structured research training programmes. The development of massive open online courses (MOOCs) and bidirectional interactive virtual spaces permit multidisciplinary partnerships between students, faculty and mentors across institutions and countries. These provide new practical opportunities for bidirectional training, presentations and classroom-based discussions around the world (either as live or recorded sessions).

The Internet allows research, clinical training and supervision to take place across the globe, and although the content and quality of the online material is important, the effectiveness of the supervision depends on the quality of the input, and learning ultimately rests on the ability and motivation of the trainee. Internet interventions have the potential to reduce manpower requirements, but without sufficient support, completion rates remain unacceptably low. There is a need to rigorously evaluate the use of these technologies in brain-disorders research training to ensure they are effective, acceptable and culturally relevant.

### Ethical challenges

The field of neuroethics is a component of bioethics that deals with the investigation, treatment and research procedures that involve the human brain and brain science. The International Neuroethics Society (<http://www.neuroethicssociety.org>) was started to promote research that would benefit people around the world.

Although all research training should include human subject research ethics, teams that focus on brain-disorders research face unique ethical challenges. People with neurological disorders are vulnerable, sometimes cognitively or physically challenged and often stigmatized, which creates special challenges when designing protocols that ensure ethical informed consent. It is essential to address these special

## BOX 4 | BIOETHICS TRAINING RESOURCES

Resources for international bioethics research training curricula can be found through the Fogarty International Centre (Fogarty) International Research Ethics Education and Curriculum Development Award (or bioethics) programme (<http://www.fic.nih.gov/ResearchTopics/Pages/Bioethics.aspx>).

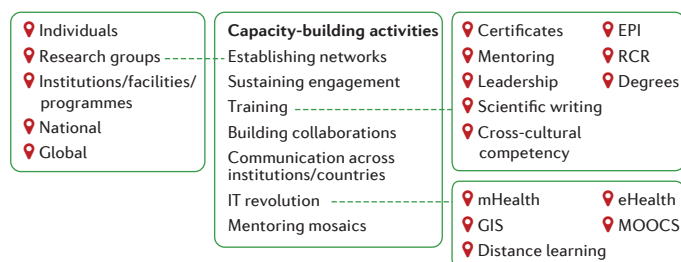
Scientist Nandini Kumar was in the first class of trainees funded by the Fogarty programme at the University of Toronto (<http://www.fic.nih.gov/News/Examples/Pages/bioethics-india.aspx>). After completing the course in 2002, Kumar was successful in her bid for a planning grant, and received a full Fogarty award to implement her programme in 2005. Over the first 7 years, Kumar's programme trained more than 2,000 scientists and health-care workers. More than 50 of them completed the intensive course and 2 earned master's degrees. The programme has hosted nearly 34 intensive workshops in 16 Indian cities. Its distance-learning programmes train up to 50 people each year. Many of the graduates have published papers, prepared curriculum for bioethics instruction at their own institutions, presented papers at national and international conferences, served as evaluators and set up or become members of ethics committees. As the field has developed, Kumar has become not only a national leader in bioethics, but also a member of international panels, including the US Presidential Commission for the Study of Bioethical Issues.

challenges (some of which are similar to those faced by researchers working with populations affected by HIV/AIDS) in training curricula for research that is specific to brain disorders. The Fogarty funded Pakistani stroke research training programme has a dedicated neuroethics training module, in which every mentored project that involves mental health research has a bioethics programme (Box 4).

Use of functional magnetic resonance imaging (fMRI), near-infrared recording systems (NIRS), polygraphy to extract information, genetic testing and cognitive enhancement using drugs and brain stimulation are just a few examples of current and evolving technologies that raise moral and ethical questions. The application of these modern techniques has been gaining momentum in the developing world, thereby indicating an urgent need to integrate neuroethics training into the neuroresearch capacity building efforts in LMICs.

## Metrics

A robust set of metrics is crucial to demonstrate the value of research capacity building to diverse stakeholders and to understand how to make research-training activities the most effective. Output measures include educational materials such as courses, modules and workshops; creation and transfer of new knowledge, such as prototypes and innovative protocols; and measuring trainee engagement indica-



**Figure 1** | Research-capacity building activities can be achieved in a number of ways. Highlighted are examples of how three of activities can be achieved. EPI, epidemiology; GIS, geographical information systems; MOOCs, massive open online courses; RCR, responsible conduct of research.

tors (for example, number of short-, medium- and long-term trainees taught, number of trainees completing courses or acquiring new skills and trainee feedback). As research training programmes mature, conventional metrics are needed, such as publications, grants, awards, memberships in societies, degrees awarded and faculty appointments.

Measuring the long-term impact of building research capacity is a significant challenge. Many funding agencies have strict guidelines for tracking career successes of funded scholars for up to 15 years after training and evaluation frameworks that can be built into programmes from the beginning to ensure trackable impacts (<http://www.fic.nih.gov/About/Staff/Policy-Planning-Evaluation/Pages/evaluation-framework.aspx>). Long-term impacts of successful capacity building include cultural competency of staff and faculty; increased involvement of staff and faculty in global health brain research; the extent to which former trainees hold positions of influence in their countries; leadership of former trainees in research and research collaborations; and increased knowledge of disorders and their significance locally and internationally.

## CONCLUSIONS

Figure 1 summarizes some of the frameworks, components, pathways and tools involved in research-capacity building. As described, research-capacity building starts at the individual level. Although partnerships between high-income and LMICs are important, the goal is for research training, as well as research itself, to increasingly take place at the LMIC sites and for those sites to become research and training hubs in their own right.

Concurrent research capacity strengthening at the institutional and national levels is necessary to ensure research and career opportunities for, and the retention of, trained researchers. With respect to NMDS disorders in particular, targeted programmes provide opportunities for LMIC clinicians, faculty and trainees to gain new skills for conducting relevant research and to contribute to long-term sustainability of research conducted in LMICs (as the trainees become the trainers and attention is paid to institutional strengths and weaknesses). Although challenges exist, they can be managed and eventually reduced or overcome using principles and models learned and shared across programmes<sup>8,33</sup>. Robust evaluations of capacity-building activities with quantitative and qualitative measures should be conducted, shared and used to identify the most successful approaches and to allow iterative improvements in individual, institutional and national level NMDS research capacity.

- Ormel, J. et al. Disability and treatment of specific mental and physical disorders across the world. *Br. J. Psychiatry* **192**, 368–375 (2008).
- Murray, C. J. et al. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* **380**, 2197–2223 (2012).
- Whiteford, H. A. et al. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *Lancet* **382**, 1575–1586 (2013).
- Global Burden of Disease 2013 Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* **386**, 743–800 (2015).
- Chin, J. H. & Vora, N. The global burden of neurologic diseases. *Neurology* **83**, 349–351 (2014).
- Jamison, D. T. et al (eds). *Disease Control Priorities in Developing Countries*, 2nd edn (World Bank, 2006).
- World Health Organization. *Mental Health Atlas 2011* (WHO, 2011).
- World Health Organization. *Neurologic Disorders: Public Health Challenges* (WHO, 2006).
- World Health Organization. *The WHO Strategy on Research for Health* (WHO, 2012).
- ESSENCE on Health Research. *ESSENCE good practice document series 31* (WHO, 2014).
- Pang, T. et al. Knowledge for better health: a conceptual framework and foundation for health research systems. *Bull. World Health Organ.* **81**, 815–820 (2003).
- Lansang, M. A. & Dennis, R. Building capacity in health research in the developing world. *Bull. World Health Organ.* **82**, 764–770 (2004).
- Thornicroft, G., Cooper, S., Bortel, T. V., Kakuma, R. & Lund, C. Capacity building in global mental health research. *Harv. Rev. Psychiatry* **20**, 13–24 (2012).

14. Toma, J. D. *Building Organizational Capacity: Strategic Management in Higher Education* 280 (Johns Hopkins Univ. Press, 2010).
15. Smith, D. G. Building institutional capacity for diversity and inclusion in academic medicine. *Acad. Med.* **87**, 1511–1515 (2012).
16. Luabeya, M. K., Mwanza, J. C., Mukendi, K. M. & Tshala-Katumbay, D. APRONES: neurology research and education in the Democratic Republic of the Congo. *Neurology* **80**, 1806–1807 (2013).
17. Uthman, O. A. *et al.* Increasing the value of health research in the WHO African Region beyond 2015 – reflecting on the past, celebrating the present and building the future: a bibliometric analysis. *BMJ Open* **5**, e006340 (2015).
18. Task force on the principles of community engagement. *Principles of Community Engagement*. 2nd edn 197 (NIH, CDC, Agency for Toxic Substances and Disease Registry, 2011).
19. Cottler, L. B. *et al.* Prevalence of drug and alcohol use in urban Afghanistan: epidemiological data from the Afghanistan National Urban Drug Use Study (ANUDUS). *Lancet Glob. Health* **2**, e592–600 (2014).
20. Katikireddi, S. V. HINARI: bridging the global information divide. *Br. Med. J.* **328**, 1190–1193 (2004).
21. Ruktanonchai, C. W., Pindolia, D. K., Striley, C. W., Odedina, F. T. & Cottler, L. B. Utilizing spatial statistics to identify cancer hot spots: a surveillance strategy to inform community-engaged outreach efforts. *Int. J. Health Geogr.* **13**, 39 (2014).
22. Pinkerton, R. C. *et al.* Evidence for genetic susceptibility to developing early childhood diarrhea among shantytown children living in northeastern Brazil. *Am. J. Trop. Med. Hyg.* **85**, 893–896 (2011).
23. Pew Research Centre. *Emerging Nations Embrace Internet, Mobile Technology* (Pew Research Center, 2014).
24. Davies, C. A., Spence, J. C., Vandelandotte, C., Caperchione, C. M. & Mummery, W. K. Meta-analysis of internet-delivered interventions to increase physical activity levels. *Int. J. Behav. Nutr. Phys. Act.* **9**, 52 (2012).
25. Martinez-Perez, B., de la Torre-Diez, I. & Lopez-Coronado, M. Mobile health applications for the most prevalent conditions by the World Health Organization: review and analysis. *J. Med. Internet Res.* **15**, e120 (2013).
26. Brouwer, W. *et al.* Which intervention characteristics are related to more exposure to internet-delivered healthy lifestyle promotion interventions? A systematic review. *J. Med. Internet Res.* **13**, e2 (2011).
27. Stinson, J., Wilson, R., Gill, N., Yamada, J. & Holt, J. A systematic review of internet-based self-management interventions for youth with health conditions. *J. Pediatr. Psychol.* **34**, 495–510 (2009).
28. Dolemeier, R., Tietjen, A., Kersting, A. & Wagner, B. Internet-based interventions for eating disorders in adults: a systematic review. *BMC Psychiatry* **13**, 207 (2013).
29. Aardoom, J. J., Dingemans, A. E., Spinhoven, P. & Van Furth, E. F. Treating eating disorders over the internet: a systematic review and future research directions. *Int. J. Eat Disord.* **46**, 539–552 (2013).
30. van Beugen, S. *et al.* Internet-based cognitive behavioral therapy for patients with chronic somatic conditions: a meta-analytic review. *J. Med. Internet Res.* **16**, e88 (2014).
31. Crutzen, R. *et al.* Strategies to facilitate exposure to internet-delivered health behavior change interventions aimed at adolescents or young adults: a systematic review. *Health Educ. Behav.* **38**, 49–62 (2011).
32. Kahn, J. G., Yang, J. S. & Kahn, J. S. 'Mobile' health needs and opportunities in developing countries. *Health Aff.* **29**, 252–258 (2010).
33. Ogundahunsi, O. A. *et al.* Strengthening research capacity — TDR's evolving experience in low- and middle-income countries. *PLoS Negl. Trop. Dis.* **9**, e3380 (2015).
34. Nuyens, Y. Setting priorities for health research: lessons from low and middle income countries. *Bull. World Health Org.* **85**, 319–321 (2007).

#### SUPPLEMENTARY MATERIAL

Is linked to the online version of this paper at: <http://dx.doi.org/10.1038/nature16037>

#### ACKNOWLEDGEMENTS

The authors were supported in part by the following grants: FIC/NIMH/NIH D43TW009120, Indo-US Training Program on Non-Communicable Diseases Across the Lifespan; FIC/NIMH/NIH D43TW009089, Increasing Mental Health Research Infrastructure in Southeast Asia; and FIC/NINDS/NIH D43TW008660, International Cerebrovascular Translational Clinical Research Training Program. Research reported in this publication was supported by the Fogarty International Center and the National Institute on Drug Abuse of the National Institutes of Health under Award Number D43 TW009120. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

#### COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests. Financial support for publication has been provided by the Fogarty International Center.

#### ADDITIONAL INFORMATION



This work is licensed under the Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0>